doi:10.5128/LV26.07

Õppijasõbralik korpuslause: automaatse valiku võimalusi

KRISTINA KOPPEL[1,2], JELENA KALLAS[1]
Eesti Keele Instituut[1], Tartu Ülikool[2]

Ülevaade. Artiklis tutvustatakse võimalusi, kuidas korpusest automaatselt tuvastada keeleõppeks sobivaid korpuslauseid. Liiga pikkade ning keerulise süntaktilise struktuuri ja leksikaalse koosseisuga korpuslausete lugemine võib keeleõppijale olla raske ning vähendada soovi keele õppimisel korpusmaterjalile toetuda. Samas on tänapäeval olemas meetodid, mis võimaldavad muuta korpuspõhise keeleõppe tõhusamaks ja huvitavamaks.

Artiklis käsitletakse esmalt, milleks ja kuidas saab korpuslauseid kasutada keeleõppes ja õppesõnastike koostamisel. Teiseks kirjeldatakse meetodeid, mis võimaldavad õppijasõbralike korpuslausete automaatset valikut. Uurimuse keskmes on reeglipõhine meetod *Good Dictionary Example* (GDEX) (Kilgarriff jt 2008), mida on seni katsetatud inglise, soome, sloveeni, hollandi, portugali, hispaania, jaapani jt keelte peal. Kolmandaks kirjeldatakse GDEXi rakendamist eesti keelele ja tutvustatakse õppijasõbralike korpuslausete automaattuvastuse jaoks vajalikke parameetreid. Meetodi katsetamiseks loodi koostöös tarkvarafirmaga Lexical Computing Ltd. eesti keele ühendkorpuse põhjal testkorpus EstonianNC GDEX, mis sisaldab ainult väljatöötatud parameetritele vastavaid lauseid. Korpus on kättesaadav korpuspäringutarkvara Sketch Engine kaudu. Tegemist on esimese katsega luua õppeotstarbeline autentseid korpuslauseid sisaldav korpus, mida saaks pakkuda otse eesti keele õppijatele. Artiklis analüüsitakse, milliseid parameetreid oleks vaja lisaks testida.

Võtmesõnad: korpuslingvistika; korpusleksikograafia; õppeleksikograafia; keeleõpe; eesti keel

1. Sissejuhatus

Tänapäeval kasutatakse keeleõppes ja keeleõppematerjalide koostamisel aina rohkem loomulikku keelekasutust illustreerivaid keelekorpusi, mis on algselt loodud eelkõige keeleuurijate vajadusi silmas pidades. Siinse artikli uurimisobjekt on eesti keele kirjalike tekstide korpuste laused ehk korpuslaused, mis vastavad "Eesti keele käsiraamatu" (EKK 2007: 429–430) järgi ortograafilise lause nõuetele. See tähendab, et laused algavad suure tähega ja lõpevad lauselõpumärgiga. Õppijasõbralikuks korpuslauseks peavad autorid autentset lauset, mis on oma grammatilise ja süntaktilise keerukuse ning leksikaalse koosseisu poolest keeleõppijale jõukohane.

Korpuste ja korpuslausete keeleõppes kasutamise küsimusi on arutatud alates 1990ndatest, kui Tim Johnsi (1991) eksperimentide tulemused näitasid, et korpuspõhine õpe on õpilastele meelepärasem kui õpikuid ja grammatikaid kasutavad traditsioonilised meetodid. Sama tulemuseni jõudsid Geoffrey Leech (1997: 12–13), Guy Aston (1997: 54–57) ja Laura Gavioli (1997: 92), kelle arvates stimuleerib korpusainesel põhinev õppimine õpilasi, esitab neile suuremaid väljakutseid, tekitab uudishimu ja mõjub motiveerivalt. Bill Doddi (1997: 131–136) sõnul on korpuspõhine õpe efektiivne viis tutvuda grammatikaga ning suurendada oma sõnavara.

Ka hiljaaegu tehtud uurimused (Frankenberg-Garcia 2012, 2014) toetavad korpuste kasutamist keeleõppes. Õppijad teevad korpusmaterjaliga töötades ise keele kohta järeldusi. Laused, mis sisaldavad vihjeid konteksti kohta, aitavad mõista uute sõnade tähendust, ning laused, mis sisaldavad kollokatsioone ja esindavad süntaktilisi mustreid, aitavad ennetada vigu, mida teist keelt õppides tüüpiliselt tehakse. Samuti on selgunud, et nii teksti mõistmise kui ka loomise puhul on õppijal rohkem abi sellest, kui ühe kasutusmustri kohta on näiteid mitu, sest lause võib sageli jääda õppija ainsaks kokkupuuteks uue sõnavara ja grammatikaga ning paikapidavate järelduste tegemiseks ühe lause esitamisest ei piisa. (Frankenberg-Garcia 2014: 130)

Teisalt on Adam Kilgarriff (2009), Elena Volodina (2008), Laura Gavioli (2005) ja James Wilson (2013) tõdenud, et kuigi korpuslingvistika pakub keele õppimiseks ja õpetamiseks palju võimalusi, pole korpused keeleõppes nii laialdaselt levinud kui võiksid. Kilgarriffi (2009: 5) arvates ei kasuta õpetajad korpuseid klassiruumis seetõttu, et need ei paista esmapilgul keeleõppija vajadusi rahuldavat. Sõnaotsing korpusest võib nõuda palju tööd, tuua kaasa palju müra ega pruugi alati õiget vastust anda.

Tõepoolest, kuidas peaks keeleõppija hakkama saama näiteks näidetes 1 (otsisõna on *raha*) ja 2 (otsisõna on *tegu*) esitatud korpuslausetega?

(1) Arvatavasti ta ta ajab seda juttu selleks, et kuna need topised maksavad väga palju ja kuna inimesel on **raha** vaja, siis ta teab, et kunagi sai Alvariga hästi läbi, tuli, oli ja nüüd läks ja võttis need ära, et saada mingit **raha**. (etTenTen)

(2) Mina ainult poleks mingeid vidinaid peale maalinud sest ebaõnnestumise korral oleks raudselt leidunud keegi, kes leidnuks, et **tegu** on väga võimsate vril-jõu sümbolitega, mis joonistati sinna telekineesivõimete neutraliseerimiseks. (etTenTen)

Sellised laused ei ole õppijasõbralikud ei sõnavara ega grammatilise keerukuse poolest. Laused sisaldavad haruldast sõnavara, esineb kordusi, kõnekeelseid väljendeid, pärisnimesid, lühendeid jmt.

Eesti keele korpuslausete formaalsete näitajate (lause keskmine pikkus, sõna keskmine pikkus, kõrvallausetega lausete arv) katseuurimus (Kallas jt 2015: 86–89) näitas, et ühes korpuslauses võib olla kuni 56 sõna, samas on sõnaraamatutes esitatud näitelausete maksimaalne sõnade arv 13. Märkimisväärne on ka kõrvallausete osakaal korpuslausetes. Näiteks lausetes, kus märksõnaks oli verb või adverb, moodustasid kõrvallausetega laused koguni 76% kogu valimist. Sõnastikes olid samas ülekaalus lihtlaused.

Selleks, et korpus ei väljastaks päringu tulemusena pikki ning süntaktiliselt ja grammatiliselt keerulisi lauseid, mis ei ole jõukohased, ja mille lugemiseks võib leksikograafil või keeleõppijal kuluda väga palju aega, on tekkinud vajadus uute meetodite järele. Need meetodid aitavad

korpuslauseid filtreerida, eristades keeleõppija jaoks sobivaid ja mittesobivaid lauseid.

2. Korpuslaused keeleõppe- ja sõnastikeportaalide osana

Tänapäeva keeletehnoloogiad võimaldavad keeleõppijal otsest juurdepääsu korpuslausetele kahel viisil:

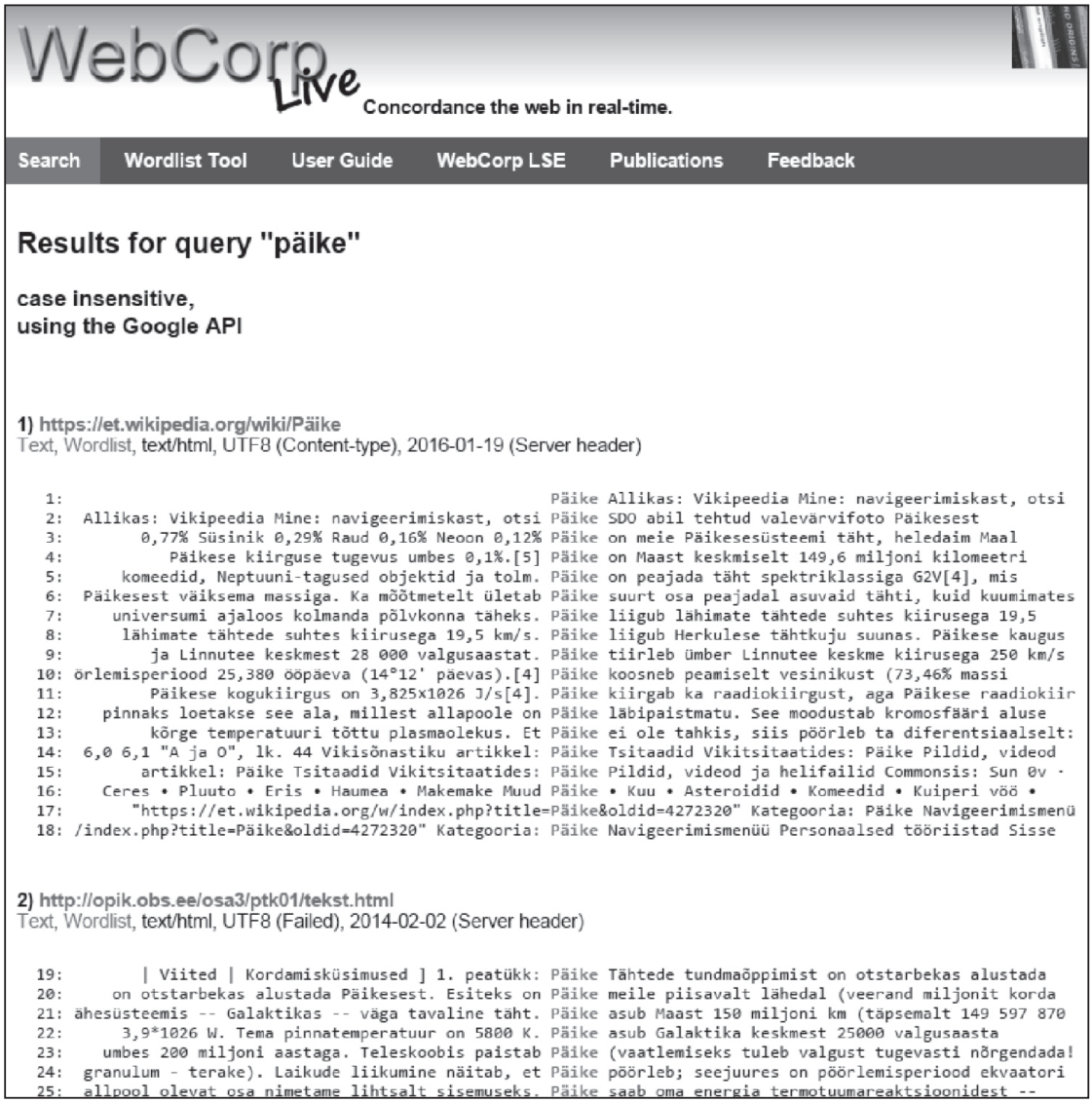
1) otsingumootorite vahendusel, nt WebCorp[1] (Kehoe & Renouf 2002);
2) korpuspäringusüsteemide vahendusel, nt IntelliText, Wordsmith Tools, Keeleveeb, KARP, SketchEngine.

Otsingumootor WebCorp käsitleb korpusena kogu veebi. Nagu tavalised otsingumootorid (nt Google, Bing) otsib programm päringule vastuseid, kuid esitab tulemused süstematiseeritult internetiaadresside põhiselt ja konkordantsiridade kujul (joonis 1).

Jooniselt 1 on näha sõna *päike* konkordantsiridu, mis on genereeritud WebCorp süsteemis. Esimesed 18 rida on pärit Wikipediast, järgmised interaktiivsest kosmoloogiaõpikust www.obs.ee. Kokku väljastas süsteem 432 rida.

Otsingumootorite vahendusel saadud päringutulemus on oma olemuselt dünaamiline, reaalajas muutuv ja uuenev. Kasutajal on võimalus vajadusel vaadata infot otseallikast. Keeleõppijale võib aga otsingutulemuste info- ja kontekstirohkus ning lausetes sisalduv müra (palju lühendeid, pärisnimesid, numbreid, sümboleid, meiliaadresse, internetiaadresse jmt) raskusi valmistada. Wilson (2013: 12) osutab, et veebi kasutamine korpusena saab lähitulevikus tavapraktikaks nii korpuslingvistikas kui ka korpuspõhises keeleõppes. See asjaolu viitab taas vajadusele välja töötada meetodid, mis võimaldavad eristada keeleõppeks sobivaid ja sobimatuid lauseid.

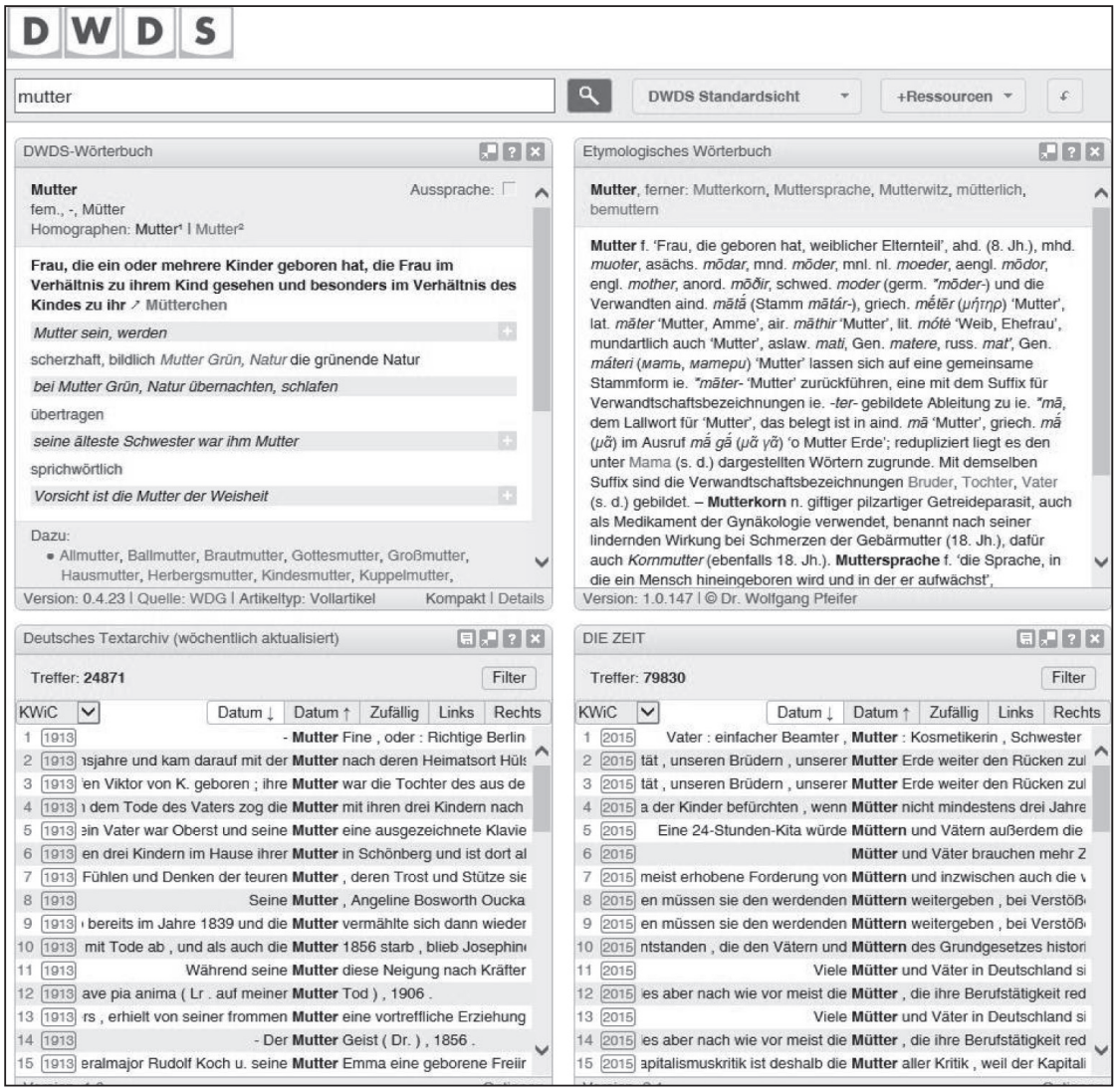
[1] Süsteem võimaldab teostada otsingut ka eesti, soome ja ungari keele jaoks. Kokku on keeli rohkem kui 30. http://www.webcorp.org.uk/live/ (4.2.2016).



Joonis 1. *Sõna* päike *konkordantsiread otsingumootorisüsteemis WebCorp*

Korpuspäringusüsteem võib olla omaette liides või sõnastikeportaali integreeritud osa. Viimasel juhul teeb keelõppija ühe päringu, mille vastuses on korraga näha nii sõna esinemine sõnastikes kui ka korpustes (joonis 2). Joonis 2 illustreerib korpuslausete esitust saksa sõnastikeportaalis Das Digitale Wörterbuch der deutschen Sprache (DWDS). Otsingutulemusena väljastatakse erinevad kastid nii sõnastiku infoga (nt etümoloogia) kui ka korpusinfoga. Alumistes kastides esitatud laused on

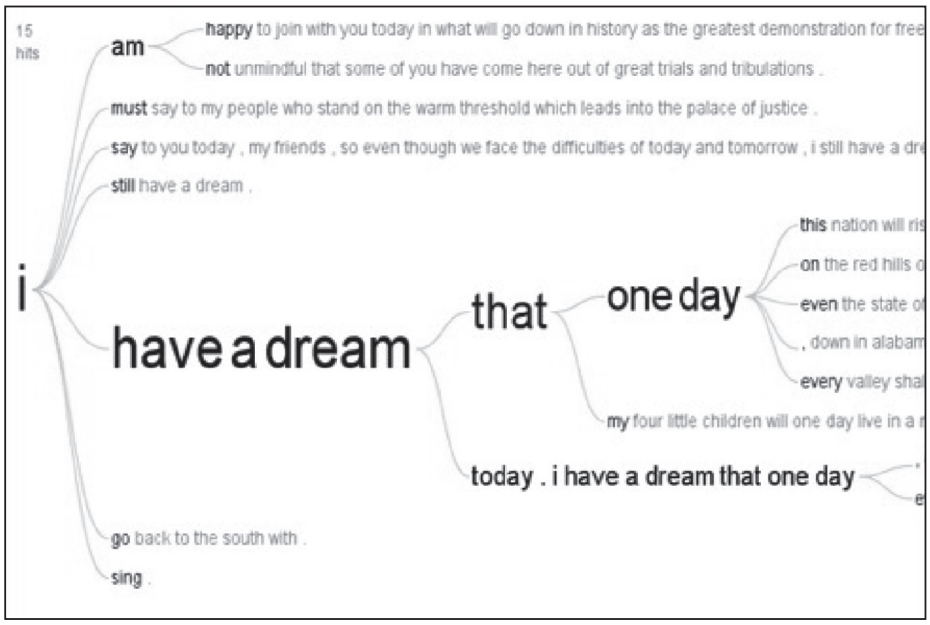
näiteks pärit korpustest Deutsches Textarchiv (1600.–1900. aastate tekstid) ja Die Zeit (XXI sajandi alguse tekstid). Eri ajastu tekste sisaldavad korpused võimaldavad võrrelda eri ajastute keelekasutust.



Joonis 2. *Sõnapäring* Mutter *'ema' DWDS sõnastikeportaalis*

Lisaks traditsioonilistele konkordantsiridadele rakendatakse korpuslausete esitamisel erinevaid visualiseerimistehnikaid. Üks sellistest tööriistadest on nt Word Tree (vt lähemalt Wattenberg & Viégas 2008), mis aitab struktureerimata teksti (nt raamat, artikkel, luuletus) lausetest korduvaid kontekste leida. Selleks grupeeritakse laused hargmiku kujuliselt.

Harud hargnevad seni, kuni jõuavad unikaalse fraasini, mis esineb tekstis ainult üks kord (joonis 3).



Joonis 3. *Teksti visualiseerimistööriista Word Tree väljund*

Vahekokkuvõtteks võib öelda, et kirjeldatud rakendustes esitatud konkordantside omapära seisneb selles, et need on genereeritud korpustest, mis ei ole koostatud spetsiaalselt keeleõppija vajadusi ja oskusi silmas pidades. Korpused koosnevad autentsetest tekstidest ja enamasti ei ole nende loomisel arvestatud keeleõppe eesmärke. Keeleõppe seisukohalt peaks ideaaljuhul olema tegemist nn pedagoogilise korpusega (Wilson 2013: 34). Selline korpus peab andma piisava ülevaate eri registrite keelekasutusest. Lisaks ei kasutata otsingumootorites ja korpuspäringusüsteemides mingeid filtreid, mis kontrolliks lausete süntaktilist keerukust ja sõnavara, ega piirata väljastatava info mahtu. Näiteks kui päringus on üle 50 lause, siis võib sellise infohulga läbitöötamine keeleõppijale üle jõu käia.

3. Korpuslaused õppeleksikograafias

Õppeleksikograafias on korpuslaused põhiline näitelausete allikas. Sõnastikku koostades võib leksikograaf võtta korpuslause kas muutmata kujul või seda vastavalt vajadusele kohandada. Tüüpilised kohandamise strateegiad on lause lühendamine, sõnade väljavahetamine, grammatiliste struktuuride lihtsustamine, pärisnimede ja lühendite asendamine. Atkins ja Rundell (2008: 458) rõhutavad, et näitelause, olgu ta autentne või kohandatud, peab vastama kolmele põhikriteeriumile: 1) loomulik ja tüüpiline; 2) informatiivne; 3) arusaadav. Hea näitelause ei tohi anda liiga palju ega liiga vähe konteksti.

Korpusleksikograafia arenguga hakati sõnastikke koostama pool- ja täisautomaatselt ning tänu sellele muutus sõnastike koostamise protsess palju kiiremaks. Automaatse koostamise meetod eeldab korpuspäringusüsteemi ja sõnastikusüsteemi olemasolu ning vastavat programmi, mille abil sõnastiku sisu, nt näitelaused, automaatselt tekstikorpusest sõnastikusüsteemi üle kanda. Üks selliseid korpuspäringusüsteeme on Sketch Engine (Kilgarriff jt 2004). Sketch Engine'i abil saab statistikapõhiselt leida kollokatsioone, koostada sõna- ja sagedusloendeid, genereerida sõna süntaktilist ja kollokatiivset käitumist illustreerivaid sõnavisandeid, koostada tesaurusi, aga valida ka näitelauseid. Heade näitelausete valimise tarbeks töötati välja spetsiaalne reeglipõhine meetod, mille nimi on *Good Dictionary Example* ehk GDEX (Kilgarriff jt 2008). Meetodi loomine oli ajendatud esialgu eelkõige leksikograafide vajadustest. Eesmärgiks oli vähendada leksikograafide ajakulu ja aidata arvutil n-ö eeltööd teha, et see valiks kõikidest korpuslausetest välja need, mis sobivad oma struktuuri ja sisu poolest leksikograafiliseks analüüsiks kõige paremini. Ideaaljuhul pidi see olema lause, mida saaks otse autentsel kujul sõnastikus esitada, kuid eeldati ka seda, et need laused võivad olla heaks aluseks n-ö kohandatud näitelausete koostamisel. Hiljem hakati seda meetodit rakendama ka laiemalt, võttes arvesse mitte ainult keeleteadlaste ja leksikograafide, vaid ka keeleõppija vajadusi (vt lähemalt ptk 4).

Sõnastiku andmebaasi täisautomaatset genereerimist, kus ka näitelaused on võetud automaatselt tekstikorpusest, on rakendatud näiteks sloveeni keele leksikograafilise andmebaasi Slovene Lexical Database (Kosem jt 2013) ja inglise keele leksikaalse andmebaasi DANTE genereerimisel. Poolautomaatset koostamist, kus vaheetapina teeb leksikograaf oma valiku arvuti välja valitud lausetest, on kasutatud näiteks sõnaraamatu "Macmillan Collocations Dictionary for Learners of English" (Rundell 2012) ja suure hollandi keele sõnaraamatu "Algemeen Nederlands Woordenboek" (Tiberius & Schoonheim 2014) koostamisel. Eestis on sõnastiku poolautomaatset koostamist katsetatud eesti kollokatsioonisõnastiku projekti (Kallas jt 2015) raames. Esimese etapina genereeriti täisautomaatselt sõnastiku andmebaas, mida leksikograafid teise etapina käsitsi puhastavad ja täiendavad.

Vaatamata sellele, et suurem osa viimasel ajal ilmunud sõnaraamatutest on koostatud korpuspõhiselt, kasutatakse neis lauseid autentsel kujul üsna vähe. Leksikograafide sõnul on lauset, mis autentsena sõnastikku sobiks, raske leida. Sellele probleemile osutavad nt soome keeleõppeportaali ConLexis koostajad (Jantunen jt 2013: 106), kes tõdevad, et sõnastiku koostamisprotsessi saaks keeletehnoloogia abil kiirendada, lastes osa tööst ära teha arvutitel.

Eelneva põhjal võib väita, et korpuslausete parameetrite uurimine on vajalik nii keeleõpperakenduste kui ka õppeleksikograafia seisukohalt.

4. Õppijasõbralike korpuslausete automaatne valik

Jörg Didakowski jt (2012: 345) on öelnud, et selleks, et näitelauseid korpusest automaatselt ekstraheerida, peaks tarkvara käituma nagu leksikograaf. Seda ülesannet tuleb masina jaoks lihtsustada, andes sellele ette reeglid, mis aitavad sõnakasutuse, lause loetavuse ja keerukuse kriteeriumitele keskendudes lauseid valida. Keeleõppeks sobilike näitelausete valikul rakendatakse kahte meetodit: 1) reeglipõhine meetod (Kilgarriff jt 2008) ja 2) masinõppe meetod (Lemnitzer jt 2015).

Siinses peatükis keskendutakse reeglipõhisele lähenemisele ja kirjeldatakse seda meetodi Good Dictionary Example ehk GDEX (Kilgarriff jt 2008) näitel. GDEX on Sketch Engine'i integreeritud funktsioon ning praegu kasutatav ainult seal.

Masinõppe abil ehitatakse arvutiprogramme, mis oma kogemustest õpivad ning kogemustele toetudes end automaatselt parendavad. Masinõppes kasutatakse põhiliselt kaht tüüpi lähenemist: juhendatud (ingl *supervised learning*) ja juhendamata õppimist (*unsupervised learning*). Juhendatud õppimise puhul antakse arvutile "õiged vastused" ette, st arvutile öeldakse, kuidas midagi teha. Juhendamata õppimise puhul antakse arvutile hulk andmeid, aga ei öelda, mida täpsemalt nendega teha. See tähendab, et masin peab ise leidma suure hulga andmete seast mingisuguse struktuuri[2]. Lihtsamalt öeldes püüab masinõpe järele aimata seda, kuidas inimaju töötab. Masinõppemeetodit on kasutatud nt saksa sõnastike näitelausete valikul (Lemnitzer jt 2015) ning see andis häid tulemusi. Eesti sõnastike näitelausete peal pole masinõpet veel katsetatud.

5. Reeglipõhine meetod Good Dictionary Example ehk GDEX

Kilgarriff jt (2008: 427) on GDEXi töö iseloomu kirjeldades öelnud, et see töötab justkui sõelana, hinnates lausete süntaktilisi ja leksikaalseid tunnuseid ning sortides konkordantsiridu vastavalt sellele, kuidas need etteantud parameetritele vastavad. Tulemusena pakub tööriist korpuslausete nimekirja, mille eesotsas on paremad ja tagaotsas halvemad kandidaadid.

Korpuslausete analüüsimisel võtab GDEX arvesse ainult mõõdetavaid omadusi: sõna või lause pikkust, teatud sõnade olemasolu (nt verbid) või puudumist (nt anafoorid, vulgarismid) lauses, sõnade sagedust jmt. GDEXi loojad ja arendajad (Kilgarriff jt 2008; Michelfeit 2015)

[2] Andrew Ng videokursus Stanfordi ülikoolis https://www.coursera.org/learn/machine-learning (4.2.2016).

peavad heaks näiteks sellist lauset, mis on puhas (ei sisalda sümboleid, veebilinke jmt) ja millest on võimalik aru saada ka ilma kontekstita. Halvaks näiteks peetakse sellist lauset, mis on liiga lühike või liiga pikk; liiga spetsiifiline (sisaldab nimesid, numbreid) või ebamäärane (sisaldab pronoomeneid, anafoore); sisaldab trükivigu, slängi, valesti kirjutatud sõnu (nt *Dood, check out my nwe ride*) või vulgarisme.

Selleks et GDEX suudaks lauseid tuvastada, tuleb kirjutada konfiguratsioonifail. Joonisel 4 on esitatud näitlikustav konfiguratsioonifail inglise keele jaoks.

```
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 4. *GDEXi konfiguratsioonifail*[3]

Konfiguratsioonifail sisaldab klassifikaatoreid parameetritega, millele lause peab vastama, ja mis ütlevad näiteks, et tegemist peab olema täislausega (*is_whole_sentence*), millised sõnad (*parsnips*) või tähemärgid (*illegal_chars*, *rare_chars*) ei tohi lauses esineda jm.

Tehniliselt töötab GDEX nii, et see hindab lauset skooriga (*GDEX score*), mis jääb 0 (halvim) ja 1 (parim) vahele, ning reastab laused skoori alusel paremuse järjekorda. Skoori väärtus sõltub lause omadusi mõõtvatest klassifikaatoritest, mis jagunevad kaheks: tugevateks (*hard classifiers*) ja nõrkadeks (*soft classifiers*). Tugevate klassifikaatorite alla liigituvad sellised parameetrid nagu: tegemist on täislausega (*is_whole_sentence*), lauses ei esine keelatud tähemärke (*illegal_chars*) ega teatud sõnu (*parsnips*). Tõenäolised heade lausete kandidaadid peavad kõikidele

[3] Allikas: https://www.sketchengine.co.uk/syntax-of-gdex-configuration-files/ (4.2.2016).

nendele parameetritele vastama. Nõrkade klassifikaatorite alla kuuluvad parameetrid, mis lause headust vähem mõjutavad, nt optimaalne pikkus (*optimal_interval*), harvad märgid (*rare_chars*) ja pronoomenid (*pronouns*). Tugevad ja nõrgad klassifikaatorid moodustavad kumbki lause üldskoorist 50% (ehk 0.5 + 0.5 = 1). GDEX hindab lauseid nii, et kõigepealt kontrollib nende vastastavust tugevatele klassifikaatoritele, seejärel nõrkadele. Kui lause ei vasta kasvõi ühele parameetrile tugevate klassifikaatorite seast, saab see 0.5 asemel skooriks 0, st kaotab poole oma üldskoorist, ja lause liigub kandidaatide nimekirja tahaotsa. Kui lause ei vasta mõnele parameetrile nõrkade klassifikaatorite seast, väheneb selle skoor vastavalt konfiguratsioonifailis ette määratud protsendi võrra, mis on alati väiksem kui 50%.

5.1. GDEXi konfiguratsioonid inglise, sloveeni, hollandi ja soome keele jaoks

GDEXit kasutati esimest korda "Macmillan Collocations Dictionary for Learners of English" (2010) sõnaraamatu koostamisel. Kuna sõnastik esitab kollokatsioone ehk sageli koos esinevaid sõnu, võeti inglise keele GDEXi konfiguratsiooni loomisel arvesse eelkõige kollokatsioonide paiknemist lauses. See sisaldas järgmisi parameetreid (Kilgarriff jt 2008: 426–427):

- lause pikkus on 10–25 sõna;
- lauses esinevad ainult 17 000 sagedasema sõna hulka kuuluvad sõnad;
- lauses ei esine pronoomeneid või anafoore;
- lause algab suure tähega ning lõpeb kirjavahemärgiga;
- kollokatsioon esineb lause lõpus;
- eelistati lauseid, kus esineb kolmas kollokaat (*third collocate*). Nt peaks eesti kollokatsiooni *raamatut lugema* puhul olema eelistatud laused, mis sisaldavad kõrge esilduvusega sõna ehk kolmandat kollokaati, nagu nt *huvitav* või *põnev* (*huvitavat, põnevat raamatut lugema*).

Kõige olulisemateks parameetriteks osutusid inglise keele puhul lause pikkus ja sõna sagedus.

Inglise keele põhjal loodud GDEXi konfiguratsioon pidi olema universaalne, kuid seda sloveeni keelel testides selgus, et mõned parameetrid on keelespetsiifilised. Järeldati, et iga keele jaoks on vaja eraldi konfiguratsiooni. (Kosem jt 2011: 153)

Sloveeni keele jaoks lõid Iztok Kosem jt (2011: 153–156) mitu erinevat konfiguratsiooni, mida omavahel võrreldi. See aitas paremini vahet teha erinevate konfiguratsioonide efektiivsusel ning heade ja halbade lausete sagedaste omaduste tuvastamisel. Kõige olulisemateks osutusid sellised parameetrid nagu lause pikkus (15–35 sõna), pärisnimede ja pronoomenite puudumine ning madala sagedusega sõnade lävi (st et lauses ei tohi esineda sõnu, mille sagedus on korpuses väga madal). Selgus, et märksõna asukoht lauses on keelespetsiifiline parameeter – kui inglise keele puhul asub see lause lõpus, siis sloveeni keeles pigem lause keskosa ja lõpu vahel. Leiti, et otstarbekas on kõrvale jätta laused, kus lemma kordub või kus esinevad sümbolid. Lisaks koostati nimekiri sõnadest, mis ei tohiks näitelauses esineda (nn must nimekiri). Sinna kuulusid näiteks släng ja vulgarismid, aga ka mõned lausealgulised adverbid, mis sidusid näite sellele eelnenud lausega. (Kosem jt 2011: 154–158). Kosemi jt (2013: 38–39) hilisemas uuringus selgus, et kolmas kollokaat ehk kollokatsiooni kollokaat on ka sloveeni keele puhul väga oluline täiendav parameeter. Samuti selgus, et GDEX annab veelgi paremaid tulemusi, kui teha iga sõnaliigi jaoks eraldi konfiguratsioonid.

Hollandi keele GDEXi konfiguratsioonis määrati lausete pikkuseks 10–25 sõna. Välditi pikki sõnu, märksõna kordusi ning anafoore ja pronoomeneid. Märksõna asukoht lauses ei ole hollandi keele puhul määrav, kuid lause peab kindlasti sisaldama 1 või 2 finiitverbi. Ka hollandi keele puhul rakendati nimekirja sõnadest, mida lause ei tohi sisaldada. Samas leiti, et kui lauses esineb madala sagedusega sõna, ei tähenda see tingimata seda, et lause ei ole hea, kuid tõdeti, et see väide vajab edasist analüüsimist. (Tiberius & Kinable 2015)

2015. aastal alustati soome keele GDEXi väljatöötamist[4]. Aluseks oli eesti keele konfiguratsioon 1.2 (Kallas jt 2015). Soome keele konfiguratsioon sisaldab järgmisi parameetreid:

- lause pikkus 5–20 sõna;
- lauses ei esine sõnu, mille sagedus on väiksem kui 3;
- lause ei alga sidesõnaga;
- lauses ei esine sõnu, mis on pikemad kui 20 tähemärki;
- lause ei alga sõnadega *kuten* 'nagu', *edellä* 'esiteks, enne', *toiseksi* 'teiseks', *lisäksi* 'lisaks', *siksi* 'seepärast, seetõttu', *näin* 'nii; nõnda', *esimerkiksi* 'näiteks', *siis* 'niisiis, järelikult, seega; siis';
- lauses ei esine halbade sõnade nimekirja kuuluvaid sõnu.

Tabelis 1 on vahekokkuvõtteks esitatud ülevaade inglise, sloveeni, hollandi ja soome keele GDEXi konfiguratsioonide peamistest parameetritest. Tabelist on näha, et teatud kriteeriumid (lause peab olema täislause, lauses ei esine pronoomeneid ega anafoore) on keeleülesed, teatud keelespetsiifilised (lause pikkus, märksõna asukoht lauses).

Tabel 1. *Inglise, sloveeni, hollandi ja soome keele GDEXi peamised parameetrid*

parameeter	inglise keel	sloveeni keel	hollandi keel	soome keel
täislause	+	+	+	+
lause pikkus	10–25	15–35	10–25	5–20
pronoomenid ja anafoorid	keelata	keelata	keelata	
märksõna asukoht lauses	lõpus	keskosa ja lõpu vahel		
must nimekiri		+	+	+
sõna sageduse piirang	+	+		+
sõna pikkuse piirang		+	+	+

[4] Soome keele konfiguratsiooni autor on Tarja Heinonen (Kotimaisten kielten keskus). Tegemist on katseprojektiga, mida on plaanis edaspidi arendada. Autorid tänavad Tarja Heinoneni materjali eest.

parameeter	inglise keel	sloveeni keel	hollandi keel	soome keel
lemma kordus		keelata	keelata	
sümbolid, meiliaadessid, URLid jmt	keelata	keelata		keelata
kolmas kollokaat	+	+		
peab sisaldama verbi			+	
ei esine pärisnimesid		+		+

Tuleb rõhutada, et kõikide eelnevalt mainitud keelte GDEXi konfiguratsioonide versioonid pole lõplikud, vaid neid arendatakse pidevalt edasi. Erinevate keelte GDEXi arendajad teevad omavahel rahvusvahelise e-leksikograafia võrgustiku ENeL kaudu tihedat koostööd.

Ühe keele jaoks võib olla mitu GDEXi konfiguratsiooni, mida kasutatakse eri sihtgruppidele mõeldud sõnastike koostamisel. Näiteks kui tegemist on emakeelsele kõnelejale suunatud sõnastikuga, võivad laused olla pikemad ja sisaldada haruldasemaid sõnu. Kui tegemist on keeleõppijale mõeldud sõnastikuga, eelistatakse lühemaid lauseid ning tõstetakse madala sagedusega sõnade läve, et lausetesse satuks võimalikult vähe haruldasi sõnu.

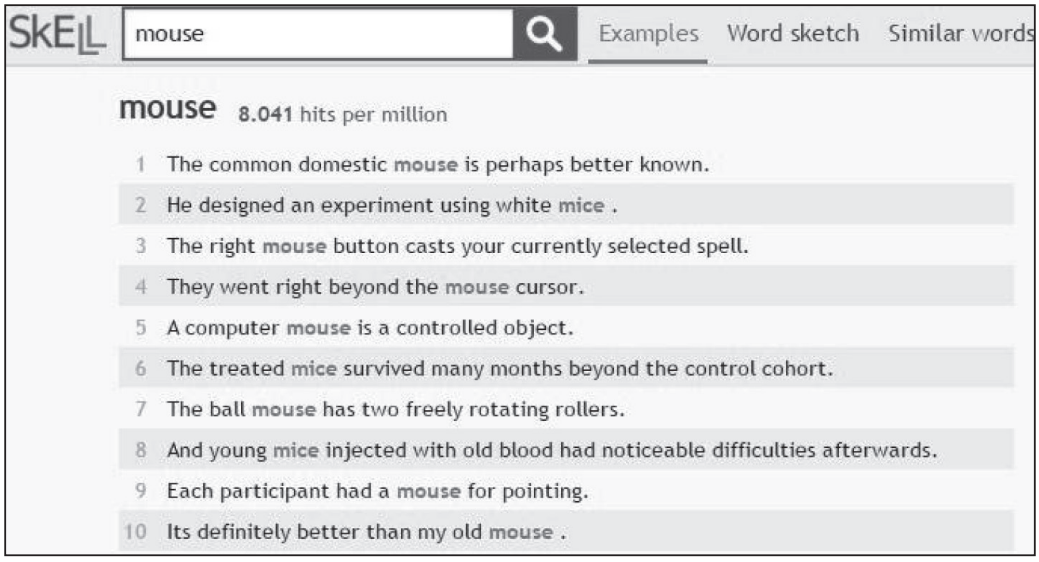
5.2. GDEXiga valitud korpuslaused keeleõpperakenduses SkELL

Tuntuim keeleõpperakendus, mille laused on välja valitud GDEXi abil, on *Sketch Engine for Language Learning* ehk SkELL (vt ka Baisa & Suchomel 2014). SkELL on keeleõpetajatele ja -õpilastele suunatud kasutajaliides, mis kasutab Sketch Engine'i erinevaid funktsioone.

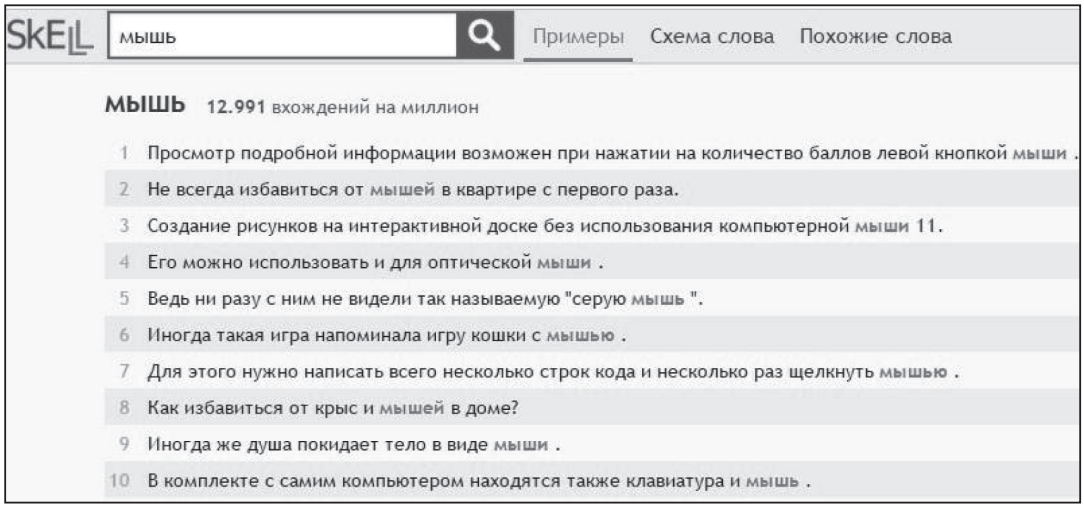
Esimesena tehti SkELL inglise keele õppijatele. Sel otstarbel loodi spetsiaalne korpus, kuhu kuuluvad uudistekstid, Wikipedia artiklid, (ilu)kirjandus, foorumid, blogid jmt. Korpuses on rohkem kui 60 miljonit lauset ja rohkem kui miljard sõna. Selline kogus tekstilisi andmeid pakub piisava ülevaate argisest, normatiivsest, formaalsest ja erialasest keelekasutusest.

Keeleõppija vajadusi silmas pidades häälestati SkELLi jaoks ümber GDEXi standardsed parameetrid. Eelistati lühikesi laused, mis sisaldavad sagedasemaid sõnu. Tauniti lauseid, mis sisaldavad keerulist terminoloogiat, haruldasi sõnu, nimesid ja kohatut keelekasutust. (Baisa & Suchomel 2014: 64)

SkELL töötab eri keelte peal ning võimaldab otsest juurdepääsu korpusmaterjalile kolmel erineval viisil. Esiteks saab keeleõppija lugeda konkordantsiridu (joonis 5 ja 6), mis illustreerivad sõna käitumist lauses.

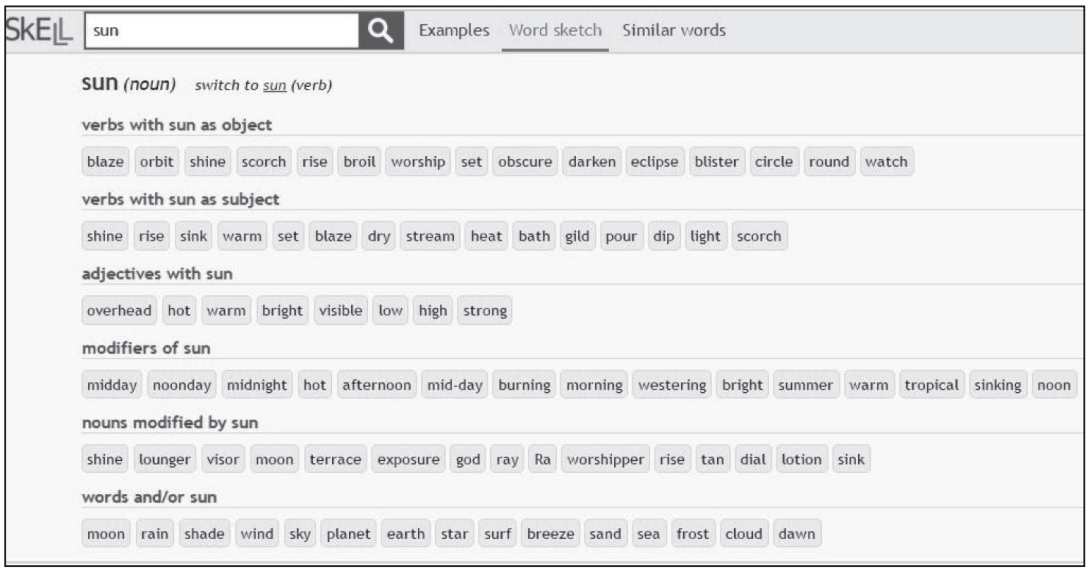


Joonis 5. *Lemma* mouse *'hiir' konkordantsiread inglise keele SkELLis*



Joonis 6. *Lemma* мышь *'hiir' konkordantsiread vene keele SkELLis*

Teiseks on SkELLi abil võimalik vaadata sõnavisandeid (*Word Sketch*). Nende abil näeb keeleõppija sõna tüüpilisemaid kollokatsioone (joonis 7) ehk sageli koos esinevaid sõnu (eesti keeles nt *ere päike*, *päike paistab*, *päikest võtma*). Kollokatsioonid illustreerivad sõna kasutust kontekstis. Kollokaadile klikkides on võimalik näitelauseid lugeda ka iga kollokatsiooni kohta eraldi.



Joonis 7. *Sõna* sun *'päike' kollokaadid inglise keele SkELLis*

Kolmandaks on veebiliidese kaudu võimalik vaadata tesaurust ehk seostuvaid sõnu, mis esitatakse sõnapilvena – kõige tugevamini seotud sõnad on suurema fondiga (joonis 8 ja 9).

Joonis 8. *Sõna* lunch *'lõuna' sõnapilv inglise keele SkELLis*

Joonis 9. *Sõna* обед *'lõuna' sõnapilv vene keele SkELLis*

SkELLi-taolisi keeleõperakendusi saab luua keeltele, mille jaoks on olemas märgendatud korpused, grammatiliste suhete tuvastamiseks koostatud sõnavisandite grammatika (*Sketch Grammar*) (Kilgarriff jt 2004) ja välja töötatud korpuslausete valiku parameetrid. Seda tüüpi keeleõppekeskkond on oluline samm korpuspõhise keeleõppe populariseerimisel. SkELLi suur eelis on kasutajaliidese tehniline lihtsus; läbipaistev metakeel (nt *thesaurus* 'tesaurus' asemel *similar words* 'sarnased sõnad'); väljastatud info piiratud maht; korpuslaused on autentsed, kuid tänu GDEXi filtrile sobivad ka algajaile keeleõppijaile.

Järgmises peatükis esitletakse katseprojektina loodud õppeotstarbelist eesti keele testkorpust EstonianNC GDEX, mille põhjal oleks võimalik luua SkELL ka eesti keele jaoks. Kokku on testkorpuses 125 790 090 sõnet, 106 168 456 sõna ja 10 599 458 lauset. Korpuse võimalik sihtgrupp on eesti keele edasijõudnud (B2-keeleoskustase) õppijad. Korpuses on 12 allkorpust: meediatekstid ja poliitilised tekstid (kuni aastani 2008 ja kuni aastani 2013), ilukirjandus, blogid, foorumid, teadustekstid, usutekstid, seadustekstid, informatiivsed tekstid, liigitamata jäänud tekstid. Päringut saab teha ka allkorpuste kaupa. Testkorpuse analüüsi tulemusena pakutakse välja, mis parameetreid on vaja lisaks testida. Analüüsi tulemused on olulised ka teiste keelte GDEXi konfiguratsioonide arendamisel.

6. GDEX eesti keele jaoks

6.1. GDEX 1.3 konfiguratsiooni parameetrid

Esimese eesti keele GDEXi konfiguratsiooni 1.2 töötas välja eesti keele kollokatsioonisõnastiku töörühm (vt lähemalt Kallas jt 2015). Kollokatsioonisõnastiku andmebaasi maht oli 10 939 märksõna, 493 971 kollokaati ja 2 469 855 korpuslauset. Iga kollokatsiooni kohta ekstraheeriti viis korpuslauset. GDEXi 1.2 analüüsi käigus tulid välja kaks põhilist probleemi:

1. Esines palju ilma öeldiseta lauseid (näide 3).

   (3) Harjumused ja rutiin, toimetulek ja eneseabi.

2. Lauses esines anafoorseid sõnu (nt proadverbe, pronoomeneid) (näited 4–6).

   (4) Seejärel võttis mõni soomlane välja rahakoti ja andis poistele raha.
   (5) Ta ei lasknud meil sinna jõuda.
   (6) Nad tõstetakse sealt välja erilise manipulaatoriga.

GDEX 1.2 konfiguratsioonis tehti vajalikud muudatused ning töötati välja GDEXi konfiguratsioon 1.3. Konfiguratsioonifaili (joonis 10) kirjutas Lexical Computing Ltd. programmeerija Jan Michelfeit detsembris 2015. Joonisel on halbade sõnade nimekiri (*bad_words*) lühemaks lõigatud.

Lisati parameeter, et lause peab kindlasti sisaldama verbi. Lausete süntaktilise lihtsuse huvides otsustati taunida *mata-*, *mast-*, *mas-*, *maks-* ja *des*-lauselühendeid sisaldavaid lauseid, kuna verbi finiitsed vormid on sageli kantseliitlikumad ja ametlikumad. Lisati nimekiri teatud sõnadest, mis ei tohi esineda lause alguses (*näiteks*, *kui*, *ühesõnaga*, *seejärel*, *nagu*), kuna enamik neist on oma olemuselt anafoorsed ehk viitavad seosele lausest välja. Anafooril on siinses artiklis laiem tähendus: siia alla kuuluvad ka deiksised ja konnektiivlaiendid. Samuti lisati parameeter, mis annab madalama skoori lausetele, kus esinevad proadverbid *siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*, *siis*.

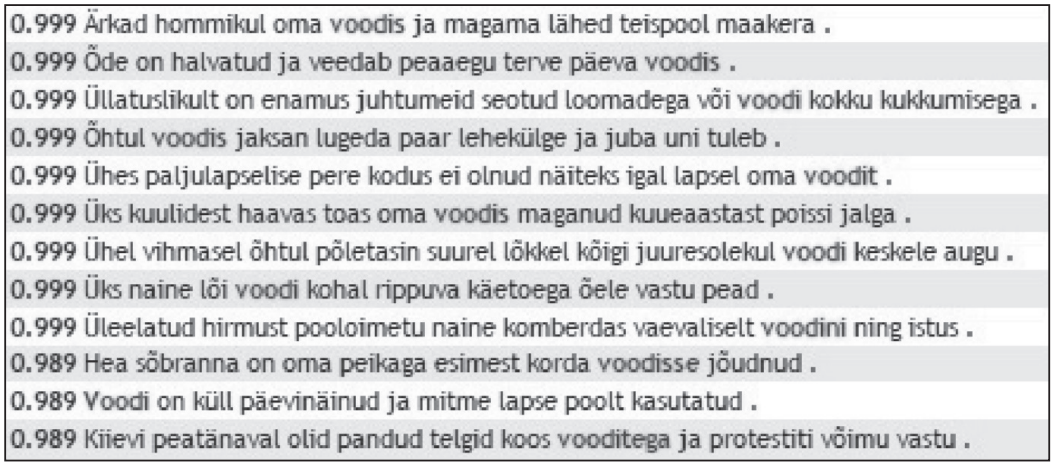
```
formula: >
   (50 * all(is_whole_sentence(), length > 5, length < 20, max([len(w) for w in words]) < 20,
count_matches(tags, verb) > 0, blacklist(words, illegal_chars), blacklist(lemmas, bad_adverbs_any),
not match(lemmas[0], bad_adverbs_first), min([word_frequency(w) for w in words]) > 5)
   + 50 * 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|sitt| homo…)$
```

Joonis 10. *Eesti keele GDEX 1.3 konfiguratsioonifail*

Eesti keele GDEX 1.3 parameetrid on järgmised:

- lause algab suure tähega ja lõpeb kirjavahemärgiga;
- lause pikkus on 5–20 sõna;
- ei esine sõnu, mis on pikemad kui 20 tähemärki;
- ei esine sõnu, mille sagedus on alla 5;
- lauses ei esine sümboleid, numbreid, pärisnimesid, lühendeid, tagasiviiteid *mina*, *sina*, *tema*, *see*, *too*, adverbe *siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*, *siis*;
- lause ei alga adverbidega *näiteks*, *kui*, *ühesõnaga*, *seejärel*, *nagu*;
- lause sisaldab verbi;
- lause ei sisalda infiniitseid verbivorme;
- lauses ei esine halbade sõnade nimekirja kuuluvaid sõnu.

Joonis 11 illustreerib konkordantside esitust testkorpuses lemma *voodi* näitel. Laused on reastatud GDEX skoori alusel, mis on nähtav lausete ees vasakus servas.



Joonis 11. *GDEX 1.3 väljund lemma* voodi *näitel*

6.2. Testkorpuse analüüs

Testkorpus EstonianNC GDEX sisaldab ainult neid lauseid, mis vastavad GDEX 1.3 konfiguratsiooni parameetritele.

Testkorpuse lausete analüüs näitas, et GDEX 1.3 tulemus on eelmistest versioonidest parem – laused on lühemad ning üldpilt ühtlasem. Väljundile mõjus efektiivselt parameeter, et lause peab sisaldama öeldist. Samas selgus, et teatud parameetrid vajavad endiselt täiustamist.

1. Täiendavalt tuleb uurida lause alguses esinevaid sõnu, mis võivad olla anafoorse tähendusega. GDEX 1.3 konfiguratsioonifailis on ainult viis sõna, mis ei ole ilmselgelt piisav. Nimekirja tuleb täiendada näiteks sõnadega *seniks* (näide 7), *muidu* (näide 8), *samuti* (näide 9), *samamoodi* jmt.

(7) Seniks aga tasub otsida abi ja viimast lootust mitte kaotada.
(8) Muidu saab salat liiga magus.
(9) Samuti ei saa tema sõnul välistada spekulante, kes kiiresti korterid kokku ostavad.

2. Tuleb uurida anafoorsete sõnade esinemist ka lause sees ja analüüsida nende kontekstisidususe määra. Sellised on näiteks personaalpronoomenid (*nemad*, *nad*) (näide 10), demonstratiivpronoomenid (*see*, *too*) (näide 11) ja proadverbid (näide 12).

(10) Selliseid näiteid esineb ja ka nemad peavad niisugustest asjadest lahti saama.
(11) Prantsusmaal on selle tarvitamiseks vajalik arsti luba.
(12) Mulle meeldib siin ja usun, et jään pikemaks ajaks.

3. Täiendavalt tuleb tähelepanu pöörata pärisnimede ja numbrite esinemisele lausetes, sest kuigi GDEX 1.3 konfiguratsioonis on määratud, et neid üksusi sisaldavad laused saavad madalama skoori, on need endiselt esil. Järgmistes konfiguratsioonides on plaanis selliseid sõnu sisaldavate lausete skoori langetada nii, et see ei väheneks mitte 5%, vaid 10% (või rohkema) võrra.

4. GDEX 1.3 konfiguratsioon sisaldab parameetrit, mis ütleb, et lauses ei esine sõnu, mille sagedus korpuses on madalam kui 5. Analüüs näitas, et see number on liiga väike, sest tihti esineb kõrge skoori saanud lausetes sõnu, mis võivad olla keeleõppijale rasked. Näiteks on näite 13 skooriks 0.989, kuigi sisaldab substantiivi *majandusmatemaatika* (sagedus ühendkorpuses 91).

(13) Teiseks on hiljutine majandusmatemaatika kiiresti edenenud arengumaade dünaamika uurimise vallas.

Näite 14 skoor on 0.962, kuigi sisaldab verbi *nääklema* (sagedus ühendkorpuses 120).

(14) Vahel on lausa hea näägelda, kuna leppimine on magus.

Edaspidi tuleb testida, kuidas tulemus muutub, kui madala sagedusega sõnade läve tõsta.

5. Tuleb testida parameetrit, mis arvestab kolmanda kollokaadi esilduvust, sest nagu Kosem jt (2013) tulemused näitasid, parandas selle parameetri lisamine tulemusi oluliselt. Nt peaks eesti kollokatsiooni *raamatut lugema* puhul olema eelistatud laused, mis sisaldavad kõrge esilduvusega sõna ehk kolmandat kollokaati, nagu nt *huvitav* või *põnev* (*huvitavat, põnevat raamatut lugema*).

6. Otstarbekas on katsetada eraldi konfiguratsioone eri sõnaliikidele ehk luua eri konfiguratsioonid substantiividele, verbidele, adjektiividele ja adverbidele.

Analüüsi kokkuvõtteks saab öelda, et GDEX 1.3 konfiguratsioon töötab, aga seda tuleb edasi arendada. Testkorpuse väärtus seisneb selles, et see on koostatud just keeleõppijaid silmas pidades – lausetest puuduvad sõnad, mille sagedus korpuses on madalam kui 5; sõnavara on kontrollitud slängi, vulgarismide, halvustavate sõnade suhtes; tegemist on täislausetega, mis sisaldavad öeldist.

Kinnitust sai John Sinclairi (1991: 13) väide, et tulemused on ainult nii head, kui hea on korpus. Eesti keele ühendkorpuse nõrkuseks on see, et domineerib ajakirjanduskeel. Ideaalis peaks korpus olema tasakaalus, st et erinevad žanrid peaksid olema võrdselt esindatud. Keeleõppe eesmärgil loodud korpuse sisu peab olema sihipäraselt valitud ja sisaldama eri näiteid nii argisest kui ka normatiivsest keelekasutusest ning arvestada tuleks ka eri registrite proportsioone.

7. Kokkuvõte

Artiklis analüüsiti korpuslausete kasutamist keeleõppes ja õppeleksikograafias. Tänapäeva keeletehnoloogiad võimaldavad keeleõppijal otsest juurdepääsu korpuslausetele kas otsingumootori või korpuspäringusüsteemide vahendusel. Korpuslauseid esitatakse enamasti konkordantsiridade kujul, mille esitamisel rakendatakse ka erinevaid visualiseerimistehnikaid.

Korpuslausete kasutamisel keeleõppes esineb mitu probleemi: laused on süntaktiliselt, leksikaalselt ja grammatiliselt liiga keerulised, esineb kordusi, kõnekeelseid väljendeid, pärisnimesid, lühendeid, slängi, vulgarisme jmt. Sel põhjusel osutus vajalikuks välja töötada meetodid, mis võimaldavad keeleõppeks sobimatute lausete tuvastamist ja kõrvaldamist.

Õppijasõbralike korpuslausete automaatsel valikul rakendatakse kahte meetodit – masinõpet ja reeglipõhist lähenemist. Mõlemaid meetodeid on edukalt katsetatud eri keelte peal. Reeglipõhist lähenemist on artiklis käsitletud meetodi *Good Dictionary Example* ehk GDEXi näitel. Artiklis kirjeldati inglise, sloveeni, hollandi ja soome keele GDEXi konfiguratsioone ning tutvustati kaasaegset keeleõppeportaali *Sketch Engine*

for Language Learning ehk SkELL, mis sisaldab GDEXi meetodiga välja valitud lauseid. Seda tüüpi portaal on oluline samm korpuspõhise keeleõppe populariseerimisel. SkELLi suur eelis on kasutajaliidese tehniline lihtsus, läbipaistev metakeel, keeleõppijale sobiv hulk korpuslauseid ja kollokatsioone, lausete grammatiline ja leksikaalne jõukohasus.

Artiklis tutvustati eesti keele GDEX konfiguratsiooni 1.3 parameetreid ning nende parameetrite põhjal loodud esimest õppeotstarbelist autentseid lauseid sisaldavat korpust EstonianNC GDEX. Eesti keele GDEXi konfiguratsiooni 1.3 loomisel arvestati eri keelte konfiguratsioone, kuid seda on täiendatud eesti keele spetsiifiliste parameetritega: lause pikkus on 5–20 sõna; sõna maksimaalne pikkus 20 tähemärki; lause ei sisalda infiniitseid verbivorme; lause ei alga adverbidega *näiteks*, *kui*, *ühesõnaga*, *nagu*; lauses ei esine sümboleid, numbreid, pärisnimesid, lühendeid, pronoomeneid *mina*, *sina*, *tema*, *see*, *too* ja proadverbe *siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*, *siis*.

Korpuse EstonianNC GDEX kogumaht on umbes 125 800 000 sõnet, 106 170 000 sõna ja 10 600 000 lauset. Korpuse väärtus seisneb selles, et see on koostatud just keeleõppijaid silmas pidades – lausetest puuduvad väga madala sagedusega sõnad; sõnavara on kontrollitud slängi, vulgarismide, halvustavate sõnade suhtes; tegemist on täislausetega, mis sisaldavad öeldist. Praegu saab keeleõppija korpusele ligi korpuspäringusüsteemi Sketch Engine kaudu. Lähitulevikus on plaanis korpus integreerida ka Eesti Keele Instituudi õppeotstarbeliste sõnastike kasutajaliidestesse.

Kirjandus

Aston, Guy 1997. Enriching the Learning Environment: Corpora in ELT. – A. Wichmann, S. Fligelstone, T. McEnery, G. Knowles (Eds.), Teaching and Language Corpora. Harlow: Longman, 51–64.

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

Baisa, Vít, Vít Suchomel 2014. SkELL: Web Interface for English Language Learning. – Eighth Workshop on Recent Advances in Slavonic Natural Language Processing. Brno: Tribun EU, 63–70.

Didakowski, Jörg, Lothar Lemnitzer, Alexander Geyken 2012. Automatic example sentence extraction for a contemporary German dictionary. – Proceedings of the 15th EURALEX International Congress. 7–11 August 2012. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo, 343–349.

Dodd, Bill 1997. Exploiting a corpus of written german for advanced language learning. – A. Wichmann, S. Fligelstone, T. McEnery, G. Knowles (Eds.), Teaching and Language Corpora. Harlow: Longman, 131–145.

EKK = Mati Erelt, Tiiu Erelt, Kristiina Ross 2007. Eesti keele käsiraamat ['Handbook of Estonian']. Tallinn: Eesti Keele Sihtasutus.

Frankenberg-Garcia, Ana 2012. Learners' use of corpus examples. – International Journal of Lexicography 25 (3), 273–296. http://dx.doi.org/10.1093/ijl/ecs011

Frankenberg-Garcia, Ana 2014. The Use of Corpus Examples for Language Comprehension and Production. – ReCall 26 (2), 128–146. http://dx.doi.org/10.1017/S0958344014000093

Gavioli, Laura 1997. Exploring texts through the concordancer: Guiding the learner. – A. Wichmann, S. Fligelstone, T. McEnery, G. Knowles (Eds.). Teaching and Language Corpora. Harlow: Longman, 83–99.

Gavioli, Laura 2005. Exploring Corpora for ESP Learners. Studies in Corpus Linguistics 21. John Benjamins Publishing. http://dx.doi.org/10.1075/scl.21

Jantunen, Jarmo Harri, Marjo Kumpulainen, Tanja Tammimies, Teemu Tokola 2013. Korpuspohjaita oppijansanakirjaa tekemässä: esimerkinä ConLexis ['Towards a corpus-based online learner dictionaly: ConLexis']. – Lähivõrdlusi. Lähivertailuja 23, 89–120. http://dx.doi.org/10.5128/LV23.04

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. University of Birmingham.

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

Kehoe, Andrew, Antoinette Renouf 2002. WebCorp: Applying the Web to linguistics and linguistics to the Web. – WWW 2002 Conference, Honolulu, Hawaii.

Kilgarriff, Adam 2009. Corpora in the classroom without scaring the students. – Proceedings of the 18th International Symposium on English Teaching, Taipei.

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

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

Kosem, Iztok, Polona Gantar, Simon Krek 2013. Automation of lexicographic work: An opportunity for both lexicographers and crowd-sourcing. – I. Kosem, J. Kallas, P. Gantar, S. Krek, M. Langemets, M. Tuulik (Eds.), Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013, 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. – I. Kosem, K. Kosem (Eds.), Electronic Lexicography in the 21st Century: New Applications for New Users. Proceedings of the eLex 2011 conference, Bled, 10–12 November 2011. Ljubljana: Trojina, Institute for Applied Slovene Studies, 151–159.

Leech, Geoffrey 1997. Teaching and language corpora: A convergence. – A. Wichmann, S. Fligelstone, T. McEnery, G. Knowles (Eds.), Teaching and Language Corpora. Harlow: Longman, 1–23.

Lemnitzer, Lothar, Christian Pölitz, Jörg Didakowski, Alexander Geyken 2015. Combining rule-based approach and machine learning in a good-example extraction task for the purpose of lexicographic work on contemporary standard German. – Proceedings of the eLex 2015 conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana–Brighton: Trojina, Institute for Applied Slovene Studies, Lexical Computing Ltd, 21–31.

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

Michelfeit, Jan 2015. GDEX in Sketch Engine. – Slaidiesitlus aadressil http://www.elexicography.eu/wp-content/uploads/2015/04/gdex_Jan_Michelfeit.pdf (4.2.2016).

Rundell, Michael 2012. How the dictionary was created? http://www.macmillandictionaries.com/features/how-dictionaries-are-written/macmillancollocations-dictionary/ (4.2.2016).

Sinclair, John 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Tiberius, Carole, Dirk Kinable 2015. Using and configuring GDEX for Dutch. – Slaidiesitlus aadressil http://www.elexicography.eu/wp-content/uploads/2015/04/ENeLWG3_GDEX4Dutch.pdf (4.2.2016).

Tiberius, Carole, Tanneke Schoonheim 2014. The Algemeen Nederlands Woordenboek (ANW) and its Lexicographical Process. – Vera Hildenbrandt (Ed.), Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie".

Volodina, Elena 2008. From corpus to language classroom: reusing Stockholm Umeå Corpus in a vocabulary exercise generator SCORVEX. University of Gothenburg.

Wattenberg, Martin, Fernanda B. Viégas 2008. The word tree, an interactive visual concordance. – IEEE Transactions on Visualization and Computer Graphics 14 (6), 1221–1228.

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). Internetis aadressil: https://www.heacademy.ac.uk/sites/default/files/corpus_technology_pedagogy_promotion2.pdf (4.2.2016).

Võrgumaterjalid

ConLexis. Veebisõnastik. http://wiki.virtues.fi/conlexis/ (4.2.2016).

DANTE. Inglise keele leksikaalne andmebaas. http://www.webdante.com/ (4.2.2016).

DWDS: Das Digitale Wörterbuch der deutschen Sprache. Sõnastikeportaal. www.dwds.de (4.2.2016).

ENeL. Euroopa e-leksikograafia võrgustik. http://www.elexicography.eu/ (4.2.2016).

etTenTen13. Eesti veebikorpus. http://www2.keeleveeb.ee/dict/corpus/ettenten/ (4.2.2016).

IntelliText. Korpuspäringusüsteem. http://corpus.leeds.ac.uk/it/ (4.2.2016).

Keeleveeb. Korpuspäringusüsteem. www.keeleveeb.ee (4.2.2016).

KARP. Korpuspäringusüsteem. http://spraakbanken.gu.se/karp/ (4.2.2016).

Machine Learning. Masinõppe videokursus Stanfordi ülikoolis. https://www.coursera.org/learn/machine-learning (4.2.2016).

SkELL. Keeleõpperakendus inglise keelele. https://skell.sketchengine.co.uk/ (4.2.2016).

SkELL. Keeleõpperakendus vene keelele. http://ruskell.sketchengine.co.uk/run.cgi/skell (4.2.2016).
Sketch Engine. Korpuspäringusüsteem. https://www.sketchengine.co.uk/ (4.2.2016).
WebCorp. Otsingumootor. http://www.webcorp.org.uk/live/ (4.2.2016).
Wordsmith Tools. Korpuspäringusüsteem. http://www.lexically.net/wordsmith/ (4.2.2016).
Word Tree. Teksti visualiseerimistööriist. http://hint.fm/projects/wordtree/ (4.2.2016).

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

Jelena Kallas
Roosikrantsi 6, 10119 Tallinn, Estonia
jelena.kallas@eki.ee

User-friendly corpus sentence: Parameters for automatic selection

KRISTINA KOPPEL[1,2], JELENA KALLAS[1]
Institute of the Estonian Language[1], University of Tartu[2]

The paper presents how corpus sentences can be used in learners' lexicography and in data-driven language learning.

There are two methods for the automatic selection of corpus sentences suitable for language learners: machine learning methods and rule-based methods. The paper focuses on the rule-based methods and describes them through the example of a tool called GDEX (Good Dictionary Example) (Kilgarriff et al. 2008). GDEX helps automatically select sentences suitable for language learners. It takes into account certain parameters: sentence and word length, threshold of low frequency words, keyword position, the absence and presence of certain words etc. The paper introduces the parameters of Estonian GDEX configuration and discusses which parameters need to be studied further.

The paper also introduces the new corpus EstonianNC GDEX, aimed at language learners. The corpus contains only sentences that meet the requirements for Estonian GDEX configuration. In the sentences there are no low frequency words, vocabulary is controlled (no slang, vulgarisms or profanities occur), and all sentences are full sentences and contain verbs. At the moment, the new corpora is accessible only in the corpus query system Sketch Engine (Kilgarriff et al. 2004). In future, it will be possible to integrate it into dictionary portals aimed at language learners.

Keywords: corpus linguistics; corpus lexicography; learners' lexicography; language learning; Estonian