doi:10.5128/ERYa13.04

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 13, 53–71

HEADE NÄITELAUSETE AUTOMAATTUVASTAMINE EESTI KEELE ÕPPESÕNASTIKE JAOKS

Kristina Koppel

Ülevaade. Artiklis keskendutakse tööriista Good Dictionary Example ehk GDEX (Kilgarriff jt 2008) eesti mooduli versiooni 1.4 loomisele. GDEX on tööriist, mis aitab sõnastiku näitelauseks sobivaid korpuslauseid automaatselt tuvastada. GDEX-i moodul on seni loodud inglise, sloveeni, hollandi, portugali, hispaania, jaapani ja eesti keele jaoks. Siinses artiklis seletatakse esmalt lahti tööriista üldised tööpõhimõtted. Seejärel keskendutakse näitelauseid tuvastavate parameetrite statistilisele analüüsile ja parameetrite väärtuste määramisele. Parameetrite väärtuste hindamisele ning eri moodulite võrdlusele toetudes pakutakse välja eesti mooduli uus versioon 1.4.*

Võtmesõnad: korpusleksikograafia, korpuslingvistika, õppeleksikograafia, keeleõpe, kollokatsioonid, näitelaused, GDEX, eesti keel

1. Sissejuhatus

Korpusleksikograafia arenguga on hakatud sõnastikke pool- ja täisautomaatselt koostama ning tänu sellele on koostamisprotsess muutunud palju kiiremaks. Kui varem otsisid leksikograafid näitelauseid käsitsi nt ilukirjandusest või tekstikorpustest, neid vajadusel redigeerides, või mõtlesid laused ise välja, siis tänapäeval kasutatakse näitelausetena üha enam korpusest pärit autentseid lauseid.

Heade näitelausete automaattuvastamine on tänapäeva korpusleksikograafia üks keskseid ülesandeid. Eesti leksikograafias tekkis võimalus näitelausete automaatseks tuvastamiseks 2014. aastal, kui Eesti Keele Instituudis genereeriti täisautomaatselt eesti keele kollokatsioonisõnastiku (KOLS) (Kallas jt 2015) andmebaas. Täisautomaatne genereerimine tähendab seda, et kogu sõnastiku sisu (märksõnad, kollokatsioonid, näitelaused) kantakse spetsiaalse rakendusprogrammi abil

* Eesti keele kollokatsioonisõnastiku koostamist toetab Haridus- ja Teadusministeeriumi riiklik programm "Eesti keel ja kultuurimälu II". Sõnastik valmib 2018. aastal. Elektrooniline käsikiri on Eesti Keele Instituudi sõnastikusüsteemis EELex. Siinse artikli põhjaks olev uurimistöö viidi läbi ISCH COST tegevuse IS10305 European Network of e-Lexicography (ENeL, www.elexicography.eu) lühiajalise teadusliku missiooni käigus, mis toimus Ljubljana ülikoolis dr Iztok Kosemi juhendamisel 12.6.–12.7.2016. Autor tänab dr Iztok Kosemi ja Lexical Computing Ltd. tarkvaraarendajat Jan Michelfeiti koostöö, Ülle Viksi ja Indrek Heina testandmebaaside loomise ning Trojina Instituudi teadurit dr Cyprian Laskowskit skripti kirjutamise eest.

korpusest otse sõnastikusüsteemi. Andmebaasi automaatsele genereerimisele järgneb leksikograafi töö – üleliigsete kollokatsioonide kustutamine ja puuduvate lisamine, aga ka ekstraheeritud korpuslausete hulgast sobivate näitelausete valimine.

2. GDEX ja selle eesti moodulid

2.1. Kuidas GDEX lauseid hindab: tugevad ja nõrgad klassifikaatorid

GDEX (Kilgarriff jt 2008) on korpuspäringusüsteemi Sketch Engine[1] (Kilgarriff jt 2004) integreeritud tööriist, mis hindab lauseid klassifikaatorite abil, mis mõõdavad lause leksikaalseid ja süntaktilisi tunnuseid (sõna ja lause pikkust, sõnade sagedust korpuses, teatud sõnade olemasolu (nt verbid) või puudumist (nt pronoomenid) lauses jmt), ja väljastab korpuslaused vastavalt sellele.

Selleks, et GDEX suudaks lauseid tuvastada, on vaja kirjutada iga keele jaoks konfiguratsioonifail, kus on kirjas kõik parameetrid, millele hea näitelause peab vastama. Siinkohal on paslik lühidalt selgitada põhilisi termineid. Parameeter (*parameter*) ütleb, millisele tingimusele hea näitelause vastab (nt *lause lõpeb kirjavahemärgiga*; *lause ei sisalda pronoomeneid*). Klassifikaator (*classifier*) on algoritm, mis analüüsib sisendiks oleva andmestiku vastavust etteantud parameetritele ning määrab kindlaks selle sobivuse (nt kui lause ei lõpe lauselõpumärgiga, tuleb vähenda lause skoori 50% võrra). Parameetreid sisaldavad klassifikaatorid pannakse kirja konfiguratsioonifaili (*configuration file*), vt joonis 1.

```
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. GDEX-i konfiguratsioonifaili näidis inglise keele jaoks[2]

Tehniliselt hindab GDEX 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 ja nõrkadeks. Tugevate klassifikaatorite (*hard classifiers*) alla liigituvad parameetrid, millele lause peab alati vastama (vt teine rida joonisel 1): tegemist on täislausega (*is_whole_sentence*), lauses ei esine keelatud kirjamärke (*illegal_chars*) ega sõnu (*parsnips*). Nõrkade klassifikaatorite (*soft classifiers*) alla kuuluvad parameetrid, mis lause skoori vähem mõjutavad (vt read 3–5 joonisel 1): lause pikkuse optimaalne vahemik (*optimal_interval*), harvade kirjamärkide (*rare_chars*) ja pronoomenite (*pronouns*) esinemine lauses.

Tugevad ja nõrgad klassifikaatorid moodustavad kumbki lause üldskoorist 50% (ehk 0.5 + 0.5 = 1). Selle lause skoor, mida heaks peetakse, jääb alati 0.5 ja

[1] https://www.sketchengine.co.uk/ (1.9.2016).
[2] https://www.sketchengine.co.uk/syntax-of-gdex-configuration-files/ (1.9.2016).

1 vahele. Tugevad klassifikaatorid on teineteisest vastastikku sõltuvad, st et lause peab vastama kõikidele tingimustele, et see saaks 50% üldskoorist. Lause, mis ei vasta kasvõi ühele parameetrile tugevate klassifikaatorite seast, kaotab automaatselt pool oma skoorist, ja see liigub kandidaatide nimekirja tahaotsa.

Nõrgad klassifikaatorid ei ole teineteisest vastastikku sõltuvad. Iga klassifikaatori skoor arvutatakse eraldi (skaalal 0–1) ning iga klassifikaator moodustab sama suure osa lause ülejäänud 50%-st. Igale parameetrile määratakse karistus (*penalty*), mille lause sellele mitte vastamise eest saab. Näiteks on joonisel 1 pronoomeneid tauniv klassifikaator *greylist(tags, pronouns, 0.1)*, kus 0.1 tähendabki pronoomeni esinemise eest lausele määratud karistust, mis vähendab lause skoori selle individuaalse klassifikaatori real 10%. Kui lauses ei ole ühtegi pronoomenit, saab lause selle konkreetse klassifikaatori real skooriks 1. Kui lauses esineb üks pronoomen, on skoor 0.9; 2 pronoomeni korral 0.8 jne. Kõik nõrgad klassifikaatorid korrutatakse 50-ga. Oletame, et GDEX leiab korpusest lause, mis vastab kõikidele nõrkadele parameetritele (antud juhul kahele: optimaalne vahemik on 10–14, lause ei sisalda harvu kirjamärke) peale ühe (lauses esineb pronoomen). Kui lause sisaldab ühte pronoomenit, saab see lause üldskooriks 0.95. Skoori taga peituv arvutuskäik on järgmine: iga klassifikaatori rea eest, millele lause vastab, annab GDEX skooriks 1, ühe pronoomeni esinemise eest annab aga 1 asemel skooriks 0.9. 0.9 korrutatakse 50-ga, mis annab tulemuseks 45. 45-le liidetakse tugevate klassifikaatorite eest veel 50, mis annab kokku 95. See moodustabki lause üldskoori (0.95). Sama arvutuskäiguga saaks kahe pronoomeniga lause üldskooriks 0.90 (50x0.8 = 40; 40+50=90), kolme pronoomeniga lause üldskooriks 0.85 (50x0.7 = 35; 35+50=85) jne.

Ühe nõrga klassifikaatori mõju üldskoorile sõltub nõrkade klassifikaatorite arvust – mida vähem nõrku klassifikaatoreid, seda suurem mõju on igal klassifikaatoril lause üldskoorile. Lause üldskoori saab veel mõjutada nii, et igale nõrgale klassifikaatorile määratakse oma kaal, kus suurema kaalu korral mõjutab vastav nähtus lause üldskoori rohkem.

Selleks, et GDEX suudaks eri keelte näitelauseid tuvastada, on vaja esmalt välja selgitada, millised on need keelespetsiifilised tunnused, mis vastava keele lauseid iseloomustavad.

2.2. GDEX eesti moodulid 1.2 ja 1.3

Eesti mooduli esimene versioon 1.2 loodi KOLS-i näitelausete ekstraheerimiseks 2014. aastal, ning seda on aegamööda edasi arendatud (Kallas jt 2015; Koppel, Kallas 2016). KOLS-i andmebaasi ekstraheeriti korpusest ca 2 500 000 lauset (5 lauset kollokatsiooni kohta) (Kallas jt 2015). Toimetamise käigus tuli välja mitmeid probleeme, millega GDEX 1.2 loomisel ei arvestatud (nt esines palju öeldiseta lauseid ning anafoore), ning mida püüti parandada versiooniga 1.3 (GDEX 1.3). (Koppel, Kallas 2016)

Keeleõppesõnastike näitelaused peavad olema võimalikult kontekstivabad, need peavad näitama kollokatsiooni tavapärast kasutust ja/või aitama tundmatu sõna tähendusest aru saada. KOLS-i koostamise käigus selgus, et mõnikord pole andmebaasi ekstraheeritud korpuslausete seast võimalik sobivat näitelauset valida. Põhjuseid on mitu: lause keeruline süntaktiline ja leksikaalne koostis ning lause pikkus. Lisaks mõjutas ka kollokatsiooni enda sagedus korpuses. Juhul kui sagedus

oli kuni 5, ekstraheeriti korpusest kõik olemasolevad laused, mille hulgas oli ka GDEX-i parameetritele mitte vastavaid lauseid. (Koppel, Kallas 2016) GDEX-i eesti mooduli väljundi parandamiseks otsustati võrrelda olemasolevaid versioone ja testida eri parameetrite häälestust, selleks et testimise käigus tehtud tähelepanekute põhjal saaks välja töötada uue versiooni.

2015. aastal loodi GDEX 1.3 parameetritele vastavatest lausetest korpus EstonianNC GDEX. See on esimene autentseid lauseid sisaldav õppeotstarbeline korpus ning on kättesaadav korpuspäringusüsteemi Sketch Engine kaudu. Korpuse lauseid analüüsides selgus, et teatud parameetrid (anafooride, pärisnimede ja numeraalide esinemine, madala sagedusega sõnade lävi) vajavad endiselt täpsustamist ja täiendamist. (Koppel, Kallas 2016) See sai versiooni GDEX 1.4 loomise ajendiks.

GDEX 1.3 ja 1.4 tulemusi võrreldi lisaks kandidaatide järjekorrale ka GDEX skoori põhjal – nii lause üldskoori kui ka iga klassifikaatori skoori põhjal.

2.3. Näitelausete testandmebaasid GDEX 1.4 jaoks

6. juunil 2016 loodi kaks testandmebaasi selleks hetkeks koostatud KOLS-i näitelausete põhjal (kokku 3893 artiklit): nn heade ja nn halbade näitelausete andmebaas. Heade näitelausete andmebaasi läksid laused, mille leksikograaf oli algsetest GDEX 1.2 abil ekstraheeritud lausetest välja valinud (kokku 44 038 lauset). Laused, mis valituks ei osutunud, liikusid halbade lausete andmebaasi (kokku 128 239 lauset).

Selleks, et andmebaase oleks võimalik analüüsida, märgendati need esmalt tööriistaga TreeTagger[3] (Schmid 1994). Seejärel määrati kindlaks, missuguseid parameetreid tahetakse lähemalt uurida (lause pikkus, märksõna asukoht lauses, lause esimese sõna sõnaliik jmt), ning viidi läbi testandmebaaside statistiline analüüs.[4]

GDEX-i eri versioonide (1.2, 1.3, 1.4) väljundeid ehk lause kandidaatide järjekorda analüüsiti eesti keele ühendkorpuse EstonianNC (ca 563 mln sõnet) põhjal. Selleks valiti välja 40 märksõna (10 substantiivi, 10 adjektiivi, 10 adverbi, 10 verbi) erinevatest sagedusklassidest (kõrge (>5000 ja rohkem), keskmise (1000–5000) ja madala sagedusega (<1000) sõnad). Iga märksõna kohta valiti välja üks grammatiline suhe (nt *Adj_modifier*, *modifies*, *object*, *V_modifies*), mille alt omakorda valiti välja kolm kollokatsiooni (nt *noor inimene*, *elav inimene*, *hea inimene*; *pidulikult tähistama*, *pidulikult avama*, *pidulikult lõpetama*). Analüüsimiseks kasutati tööriista GDEX Editor[5], kuhu on eesti keele ühendkorpus EstonianNC integreeritud. GDEX Editori loomise vajadus oli ajendatud sellest, et klassifikaatorite häälestamine ja testimine oli seni olnud GDEX-i konfiguratsioonifailide kirjutajatele üsna tülikas ülesanne – see nõudis konfiguratsioonifailide korduvat allalaadimist, toimetamist ja üleslaadimist. GDEX Editori kasutajaliidese abil saab parameetreid mugavalt häälestada ning kõik konfiguratsioonis tehtud muudatused on kasutajale kohe nähtavad. Joonisel 2 on kuvatõmmis GDEX Editori kasutajaliidesest.

[3] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/ (1.9.2016). Andmebaasid märgendas Lexical Computing Ltd. tarkvaraarendaja Jan Michelfeit.

[4] Testandmebaasid ei peegelda korpuse tegelikku sisu, kuna nad sisaldavad teatud kriteeriumite alusel (GDEX 1.2) korpusest valitud lauseid. Samuti pole kõik nn head näitelaused autentsed korpuslaused – osad on redigeeritud või leksikograafi koostatud. Ka kõik nn halvad laused ei ole tingimata halvad, lihtsalt leksikograaf on pidanud ühe (viiest) valima.

[5] GDEX Editori arendas Lexical Computing Ltd. tarkvaraarendaja Jan Michelfeit 2016. a juunis. Tulevikus on plaanis GDEX Editor integreerida korpuspäringusüsteemi Sketch Engine. https://beta.sketchengine.co.uk/gdex_editor (1.9.2016).

Joonis 2. GDEX Editori kasutajaliides

Sama andmestikku (antud juhul eesti ühendkorpust) kasutades saab GDEX Editoris kahte erinevat versiooni omavahel võrrelda. Samuti on kasutajaliidese kaudu võimalik näha lause üldskoori liigendust – see annab täpse ülevaate sellest, kuidas iga individuaalne klassifikaator näitelause üldskoori mõjutab. Need teadmised võimaldavad GDEX-i versioone hõlpsamini arendada.

Järgnevas keskendutakse GDEX 1.4 olulisimatele parameetritele, mille häälestust GDEX Editori abil testiti. Parameetreid vaadeldakse sõna- ja lausetasandil, aga ka sõnaliigi ja sõnaühendi tasandil.

3. GDEX 1.4: tugevad klassifikaatorid

Versioonis 1.4 on 12 parameetrit, millele lause peab alati vastama.[6] Olemasolevatele parameetritele (nt lause lõpeb lauselõpumärgiga; lauses ei esine väga madala sagedusega sõnu; lause sisaldab verbi) lisati mitu uut parameetrit: lause alguses keelatud sõnad, sõnapaarid ja sõnaliigid; aga ka märksõna kordus. Üht parameetrit (lause pikkus) korrigeeriti. Kõik uuendused toetuvad testandmebaaside analüüsile.

3.1. Lause pikkus

Versioonides 1.2 ja 1.3 oli lause pikkuseks määratud 5–20 sõnet (sh üks sõne on lauselõpumärk). Kuna sõnastiku kasutajat ei taheta koormata liiga pikkade lausetega, peab leksikograaf neid tihti lühendama. Laused, kus on ainult kolm sõna, ei ole

[6] GDEX-i moodulit arendatakse pidevalt edasi ning parameetreid võib tulevikus veelgi lisanduda.

eesti keeles ebaharilikud. Versioonidega 1.2 ja 1.3 kaotavad kolmesõnalised laused (näited 1–3) automaatselt 50% oma skoorist, kuna rikuvad ühte tugevat klassifikaatorit – lause pikkust –, ega satu seetõttu heade näitelausete kandidaatide hulka.

(1) Adrenaliin pulbitseb veres.
(2) Buumile järgneb krahh.
(3) Debüüt kujunes edukaks.

Testandmebaaside analüüs näitas, et lühim hea näitelause oli 4- ning pikim 38-sõneline (võrdluseks: lühim halb näitelause oli 2-sõneline ning pikim 208-sõneline). Versioonis 1.4 määrati lause pikkuseks 4–20 sõnet.

3.2. Lause alguses keelatud sõnad

Lause alguses ei tohiks esineda anafoorse tähendusega sõnad, kuna need vajavad mõistmiseks konteksti. Versioonis 1.3 on parameeter, mis keelab lause alguses adverbid *näiteks*, *ühesõnaga*, *seejärel*, *kui* ja *nagu*. Testandmebaaside analüüs näitas, et esimese kolme sõna keelamine lause alguses oli õigustatud (vt näiteid 4–6), kuid *kui* ja *nagu* keelamine ei olnud (vt näiteid 7–8).

(4) Näiteks õigust rikkuda loodusseadusi, aga sellegipoolest nautida täisväärtuslikku pereelu.
(5) Seejärel peske pesumasinas või käsitsi.
(6) Ühesõnaga töötan ja mängin, naudin tööprotsessi.
(7) Kui stress kestab liiga kaua, tekib depressioon.
(8) Nagu ansambli nimest välja võib lugeda, kuuluvad bändi viis silenahkset noormeest.

Lause alguses keelatud sõnade nimekirja täiendati oluliselt. Kokku on nimekirjas 62 sõna[7]: *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*.

Näited (9–13) illustreerivad lause alguses keelatud sõnade kontekstisidusust.

(9) Nimelt, et neid soove ja ettepanekuid tuleb tõsiselt uurida.
(10) Seetõttu võib mahukamate failide allalaadimine võtta kaua aega.
(11) Ülejäänud jäävad minu meelest keskpärasele tasemele ja erilisi naudinguid ei paku.
(12) Pigem rikka mehe eralõbu tegelda väikeste, aga huvitavate asjadega.
(13) Samuti ennustas vanarahvas mardipäeval talveks ilma.

Pronoomenite *nad*, *nemad* ja *see* keelamise üle võib ehk vaielda, kuid enamasti on neist ilma kontekstita väga raske aru saada (näited 14–17).

[7] Sõnu võib tulevikus lisanduda.

(14) Nad lausa jälitavad mind, ajavad mööda linna taga.
(15) Nemad ei suhtu sellesse õnneks ka nii suure eelarvamusega.
(16) See soodustab vereringet ja muudab naha elastsemaks.
(17) Neid oli hea panna seljakotti.

Teisi personaalpronoomeneid otsustati lause alguses mitte ära keelata, kuna need ei raskenda lausest aru saamist (näited 18–20).

(18) Ma olin kaks nädalat haiguslehel.
(19) Ta andis mulle oma visiitkaardi.
(20) Me oleme parimad semud.

3.3. Lause alguses keelatud sõnapaarid

Lause alguses ei tohiks esineda ka hulk anafoorseid sõnapaare, kuna need vajavad samuti mõistmiseks konteksti. Keelatud sõnapaare on kokku 79[8]: *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*.

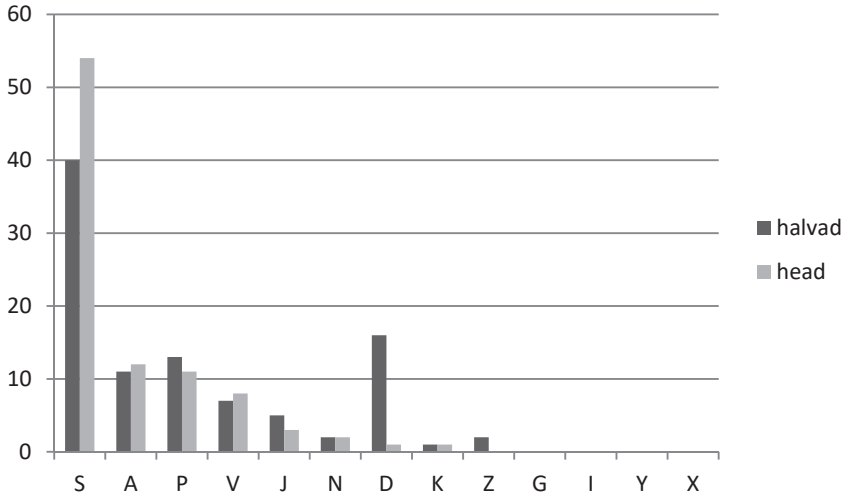
Näited (21–26) illustreerivad lause alguses keelatud sõnapaaride kontekstisidusust:

(21) Samal ajal sega kokku kastme koostisained.
(22) Eriti kui teised erakonnad ja ka meedia seda pidevalt rõhutavad.
(23) Seda enam aga hindas ja hellitas ema oma ainust poega.
(24) Sellele vaatamata on liiklusõnnetustes hukkunute ja vigastatute arv liiga suur.
(25) Just need linnad said maavärinas kõige rohkem kannatada.
(26) See tähendab, et naine on väga kaunis ja ilusa, jõulise figuuriga.

3.4. Lause esimese sõna sõnaliigiline kuuluvus

Sõnaliikide eristamise aluseks on ESTMORF-i[9] (Kaalep 1998) sõnaliikide eristus, mis lisaks traditsioonilistele sõnaliikidele (substantiiv S, verb V, adjektiiv A, pronoomen P, adverb D) eristab ka genitiivatribuuti (G, siia alla kuuluvad käändumatud omadussõnad ja kohanime vormid, nt *eesti*, *vene*, *luteri, katoliku*), lühendit (Y) ja kirjavahemärke (Z). Näitelausete esimese sõna sõnaliiki kujutab joonis 3.

[8] Sõnapaare võib tulevikus lisanduda.
[9] http://www.filosoft.ee/html_morf_et/morfoutinfo.html (1.9.2016).



Joonis 3. Lause esimese sõna sõnaliik protsentides. S – substantiiv, V – verb, A – adjektiiv, G – genitiivatribuut, P – pronoomen, D – adverb, K – kaassõna, J – konjunktsioon, N – numeraal, I – interjektsioon, Y – lühend, X – verbi juurde kuuluv sõna, mille ei ole eraldi sõnaliiki, Z – kirjavahemärk

Heade näitelausete andmebaasis ei alga laused mitte kunagi genitiivatribuudi, hüüdsõna, lühendi, kirjavahemärgi, X-ks märgendatud verbi juurde kuuluva sõna (nt *plehku*) ja tundmatu sõnaga. Seetõttu otsustati eelnevalt loetletud sõnaliigid lause alguses keelata.

Joonisel 3 esitatud info toetab eesti keele sõnajärge, kus tegevussubjekti väljendavad sõnad kipuvad esinema lause alguses. Nii head kui halvad laused algavad kõige sagedamini substantiiviga, millele järgnevad pronoomen ja adjektiiv, seejärel verb. Halbade lausete andmebaasis on adverbiga algavaid lauseid koguni 16%, samas kui heade lausete andmebaasis on neid ainult 1%. Põhjus võib peituda selles, et adverbid on tihti anafoorse tähendusega, kuid näitelaused peavad olema võimalikult kontekstivabad.

Kuigi kaassõnaga algab heade näitelausete andmebaasides ainult 1% ja numeraaliga 1–2% lausetest, otsustati neid sõnaliike lause alguses mitte ära keelata. Vastasel juhul jääksid võimalike kandidaatide nimekirjast välja nt näited 27-28:

(27) Keset udu seisab kuivanud puu, mille oksal istub vares.
(28) 1. augustiks oli kogu muuseum kolinud oma uude asupaika.

Versioonis 1.2 oli parameeter, mis ütles, et lause ei tohi alata konjunktsiooniga. Heade lausete andmebaasis esineb konjunktsiooniga algavaid lauseid aga isegi 2–3% (näited 29–30).

(29) Et pehmelt maanduda, sirutab lendorav käpad ette.
(30) Aga nüüd on õige aeg pidutsemine lõpetada ja naasta argipäeva.

Sellest hoolimata otsustati versioonis 1.4 keelata konjunktsioonidega algavad laused. See takistab selliste lausete, nagu näited (31–34), sattumist kandidaatide nimekirja etteotsa.

(31) Ja Baltikumis samal ajal analoogid puuduvad.
(32) Sest see on Eesti riigi vastutusala.
(33) Või näiteks tume rumm õunamahlaga.
(34) Ent selleks ei pea õppima usuõpetust eraldi ainena.

Enamik lauseid algab substantiiviga (vt joonis 3). Edaspidi on võimalik testida, kas selline parameeter, mis eelistab lause alguses substantiive, parandab GDEX-i väljundit.

3.5. Märksõna kordus

Sloveeni keele puhul andis märksõna kordumise keelamine paremaid tulemusi (Kosem jt 2013). Sama katsetati ka eesti keele puhul: märksõna kordus keelati, kuna näitelause eesmärk on näidata kollokatsiooni tavapärast kasutust, aga aidata ka sõna tähendusest aru saada. Märksõna kordust illustreerivad näited (35–36).

(35) Õnnelik inimene on hea inimene.
(36) Tuim ja ükskõikne inimene, kellele end pühendad, jätab ajapikku endagi tuimaks ja ükskõikseks.

4. GDEX 1.4: nõrgad klassifikaatorid

Versioonis 1.4 on 11 parameetrit, mis lause skoori vähem mõjutavad ning millele mitte vastamise eest saab lause n-ö karistada (*penalize*).[10] Osasid olemasolevaid parameetreid (nt lause pikkuse optimaalne vahemik, pronoomenite esinemine, must nimekiri) korrigeeriti, kuid lisati ka uusi (nt sõnade sagedus, lause liikmete sõnaliigiline kuuluvus ja arv lauses, nõrkade klassifikaatorite kaal).

4.1. Lause pikkuse optimaalne vahemik

Versioonides 1.2 ja 1.3 oli optimaalseks vahemikuks määratud 10–12 sõnet. See annab selle konkreetse klassifikaatori real skooriks 1. Kui lauses on kasvõi üks sõne rohkem või vähem kui parameetris määratud, saab lause juba karistada – mida kaugemal optimaalsest vahemikust, seda suurem on karistus. Järelikult peab optimaalne vahemik arvestama seda, milline on lause optimaalne pikkus (lause pikkus on 4–20 sõnet, vt ptk 3.1). Andmebaaside analüüs näitas, et hea lause keskmine pikkus on 9,83 sõnet, halva lause pikkus 14,44 sõnet.

Ka leksikograafi kogemus KOLS-i näitelausete valimisel näitab, et ekstraheeritud laused on sageli liiga pikad ja neid peab lühendama. Õppesõnastikus on eelistatud just lühemad laused, seetõttu otsustati lause optimaalset vahemikku suurendada. GDEX-i 1.4 häälestamise käigus testiti erinevaid optimaalseid vahemikke, kuid kõige paremaid tulemusi andis vahemik 6–12, mida uues versioonis ka rakendati.

Joonistel 4 ja 5 on võrdlusena näha kahe erineva versiooni väljundit.

[10] GDEX-i moodulit arendatakse pidevalt edasi ning parameetreid võib tulevikus veelgi lisanduda.

Old rank	Rank	Sentence
1	16	Lõket tohib teha ainult tuulevaikse ilmaga ning hoonetest ja metsast kaugemal .
2	9	Rannas ja tuulevaikse ilmaga sõitmiseks vajalikku mootorit mehed alles otsivad .
3	19	Suvel 30 kraadise sooja ja tuulevaikse ilmaga on probleeme kõige rohkem .
4	24	Imeilus ning suhteliselt tuulevaikne ilm lubas ka tehniliselt nauditavat ja kiiret sõitu .
5	18	Lõket võib teha ainult tuulevaikse ilmaga ning selleks ettevalmistatud kohas .
6	28	Helgi Lutvei rõhutas , et merelahtedele võib kalastama minna vaid tuulevaikse ilmaga .
7	2	Kulu on lubatud põletada päeval ja tuulevaikse ilmaga .
8	4	Soe ja tuulevaikne ilm võimaldas kalu analüüsida õues .
9	6	Lillede kolimiseks tasub valida soe ja tuulevaikne ilm .
10	23	Laev uppus teadmata põhjusel tuulevaikse ilma ja peegelsileda merepinna kiuste umbes 10 minutiga .

Joonis 4. Kollokatsiooni *tuulevaikne ilm* näitelaused optimaalse vahemikuga 10–12 (versioonid 1.2 ja 1.3)

Old rank	Rank	Sentence
37	1	Lõket tohib teha tuulevaikse ilmaga .
7	2	Kulu on lubatud põletada päeval ja tuulevaikse ilmaga .
15	3	Kulu võib põletada päeva ajal tuulevaikse ilmaga .
8	4	Soe ja tuulevaikne ilm võimaldas kalu analüüsida õues .
28	5	Juhtus olema soe ja tuulevaikne ilm .
9	6	Lillede kolimiseks tasub valida soe ja tuulevaikne ilm .
16	7	Peab ootama ilusat ja sooja tuulevaikset ilma .
14	8	Suurem tõenäosus hülgeid näha on tuulevaiksete ilmadega .
2	9	Rannas ja tuulevaikse ilmaga sõitmiseks vajalikku mootorit mehed alles otsivad .
20	10	Nad lendavad vaid niiske ja tuulevaikse ilmaga .

Joonis 5. Kollokatsiooni *tuulevaikne ilm* näitelaused optimaalse vahemikuga 6–12 (versioon 1.4)

Lühemad laused on GDEX-i väljundis eespool (*resp.* saavad kõrgema skoori), kui optimaalne vahemik on suurem (6–12).

4.2. Sõnade sagedus

Versiooni 1.3 väljundis esines sõnu, mille sagedus ei ole väga madal (nt 200 või kõrgem), kuid mis võivad keeleõppijale sellegipoolest raskusi valmistada (Koppel, Kallas 2016). Sloveeni mooduli (Kosem jt 2013) eeskujul lisati täiendav parameeter, mis taunib väikese sagedusega lemmasid. Sloveeni mooduli väljund on suunatud emakeelsele kõnelejale, siin osutus parimaks vähimaks sageduseks 200. Eesti moodul on suunatud keeleõppijale, seetõttu testiti suuremaid sagedusi, nt 400, 1000, 1200. Kõige paremad tulemused andis sagedus 1000. Kui number tõsteti

suuremaks (nt 1200), said kõrgema skoori laused, mis sisaldasid palju sagedasi abstraktseid sõnu, nagu pronoomenid ja adverbid. Sageduse parameeter aitas vähendada ka harvemaid isikunimesid sisaldavate lausete esinemist kandidaatide nimekirja eesotsas.

Joonistel 6 ja 7 on võrdlusena näha kahe erineva versiooni väljundit, millest esimeses ei ole sõnade sagedust arvestatud, kuid teises arvestatakse.

Old rank	Rank	Sentence
1	13	Suur osa saasteainetest ja kasvuhoonegaasidest satub õhku just autode heitgaasidest .
2	1	Kusjuures suures osas majast paikneb linnavalitsus ise .
3	14	Keskkonnaameti õppeprogrammid katavad suure osa loodusainete ainekavade teemadest .
4	2	Viljandis on suur tulekahju , milles hävib suur osa kesklinna puumaju .
5	3	Ta on liiga rikas , et suur osa vaesest rahvast tema poolt oleks .
6	4	Juhtimine tähendab ju suures osas planeerimist , laua taga istumist ja tulevikule mõtlemist .
7	17	Eile õhtuks olid kohalikud elanikud ja Jõhvi päästekompanii suutnud kahjutule suures osas kustutada .
8	5	Me teame , et suur osa tänapäeva rahvusvahelisest terrorismist toimub infoühiskonna meetodeid kasutades .
9	6	Suur osa sellest teadmisest on kehaline .
10	7	Suur osa rahast läks sinna .
11	18	Paljud ehitusettevõtted vahetavad katusekatte taolistel juhtudel tervikuna või suures osas mis on märgatavalt kulukam .

Joonis 6. Kollokatsiooni *suur osa* näitelaused ilma sõna sagedust arvestava klassifikaatorita (versioonid 1.2 ja 1.3)

Old rank	Rank	Sentence
2	1	Kusjuures suures osas majast paikneb linnavalitsus ise .
4	2	Viljandis on suur tulekahju , milles hävib suur osa kesklinna puumaju .
5	3	Ta on liiga rikas , et suur osa vaesest rahvast tema poolt oleks .
6	4	Juhtimine tähendab ju suures osas planeerimist , laua taga istumist ja tulevikule mõtlemist .
8	5	Me teame , et suur osa tänapäeva rahvusvahelisest terrorismist toimub infoühiskonna meetodeid kasutades .
9	6	Suur osa sellest teadmisest on kehaline .
10	7	Suur osa rahast läks sinna .
13	8	Suur osa Eestis müüdavast mööblist ongi sellest ajajärgust .
18	9	Suur osa fotodest on Stockholmi üritustest , näiteks vabariigi aastapäeva aktused , laulupidu ning Eesti koolid .
19	10	Ülemöödunud nädalal levinud uudis , et suur osa Tallinna kooliõpilastest on uimasteid proovinud , teadvustas kahtlemata probleemi tõsidust .

Joonis 7. Kollokatsiooni *suur osa* näitelaused koos sõna sagedust arvestava klassifikaatoriga (versioon 1.4)

Lühemad laused on GDEX-i väljundis eespool (*resp.* saavad kõrgema skoori), kui on arvestatud sõna sagedust.

4.3. Must nimekiri

Eesti keele ühendkorpus EstonianNC, mis oli KOLS-i lausete genereerimise aluseks, sisaldab suurel hulgal veebitekste ning sellest tulenevalt ka palju netikeelt, slängi, sõimusõnu, valesti kirjutatud sõnu jmt. Õppesõnastike puhul on andmebaasi genereerimisel oluline rakendada musta nimekirja ehk tõrjuda sõnad, mis ei sobi keeleõppijatele esitatavatesse lausetesse.

Musta nimekirja aluseks on OÜ Filosofti nimekiri sõnadest, mida eesti keele speller ei tohi valesti kirjutatud või tundmatute sõnade asenduseks pakkuda. Nimekirja täiendati sõnadega, mis on "Eesti keele seletavas sõnaraamatus" (EKSS) (2009) märgendatud stiilimärgendiga VULG, HALV, aga ka KÕNEK ja SLÄNG, kuna KOLS-i eesmärk ei ole õpetada kõnekeelt. Samuti lisati nimekirja interneti akronüüme (*omg*, *wtf*), inglise- ja venekeelseid sõimusõnu (*fuck*, *pohui*), nende mugandatud variante (*fakk*, *pohh*) ja kirjakeele normist erinevalt kirjutatud sõnu (*shantazheerima*, *zhest*). (Kallas jt 2015)

Analüüsi käigus selgus veel hulk sõnu, mis nimekirja lisati: *no*, *noh*, *aganoh*, *ika*, *kah*, *kellegil*, *kellegile*, *öigus*, *ete*, *öigustama*, *öigustus*, *pääle*, *pääl*, *päält*, *päälegi*, *türnüffel*, *bljääd*, *bljäd*, *bljää*, *sis*, *bla*, *blabla*, *blablabla*, *blablablabla*, *läits*, *lausidioot*, *täisidioot*, *idioodikari*, *pilusilm*, *kuradi*, *fashist*, *nikkuja*, *nikkumine*, *keppija*, *keppimine*, *suuseks*, *anaalseks*, *munn*, *munnikari*, *dildo*, *noks*, *pasane*, *pasapea*, *pedekas*, *autopede*, *pedepropaganda*, *pederastiapropaganda*, *pedene*, *pedendus*, *käsikiimlus*, *pornograafia*, *vitupea*, *sitamaitse*, *litsinahk*, *litsitama*, *lesbiline*, *puuksutamine*, *kakapuuks*, *perseauk*, *pissine*, *kakine*, *sitapott*, *lollike*, *crash*, *story*, *awesome*.[11]

Mustast nimekirjast eemaldati sõnad *okse*, *diiler*, *pilu*, *vänt*, *argpüks*, *hulkur*, *tumba*, *pabul*, *närukael*, *trulla*, *kanep*, *šoppama* ja *tatt*, kuna enamik neist lausetest olid neutraalsed (näited 37–39).

(37) AS United Motors on alates juunist BMW, Roveri ja Land Roveri sõidukite ametlik maaletooja ja diiler Eestis.
(38) Kanep on ka kõige enam konfiskeeritud uimasti Euroopa Liidus.
(39) Talle meeldib šopata ning telekast "Seksi ja linna" vaadata.

Tulevikus tasub musta nimekirja kuuluvad sõnad liigitada erinevatesse kategooriatesse, nt vulgarismid (*türa*, *sitt*) ja tundlikud sõnad (*tibla*, *pilusilm*, *loll*), ning jaotada need vastavalt tugevate ja nõrkade klassifikaatorite alla. Nii on võimalik vulgarismid lausest täielikult keelata, kuid tundlikke sõnu sisaldavad laused saaksid lihtsalt karistada.

4.4. Pronoomenid

Versioonid 1.2 ja 1.3 sisaldavad klassifikaatorit, mis taunib pronoomeneid *mina*, *sina*, *tema*, *see*, *too*. GDEX 1.4 häälestamise käigus testiti klassifikaatorit, mis taunib ka pronoomeneid *meie*, *teie*, *nemad*, *ma*, *sa*, *ta*, *me*, *te*, *nad*. Analüüs näitas, et isikuliste asesõnade taunimine lausetes ei ole õigustatud (v.a *nad* ja *nemad*, aga ka *see* lause alguses, vt lähemalt ptk 3.2.), sest nende esinemine ei raskenda lausest aru saamist (näited 40–41). Lisaks näitab leksikograafi kogemus, et lauseid redigeerides asendatakse isikunimed tihtipeale just pronoomenitega.

(40) Ta tahab varjata oma kommunistlikku minevikku praeguse valitsuse ees.
(41) Kuigi meie traditsioonid on saanud alguse kauges minevikus, ei ole nad püsinud muutumatutena.

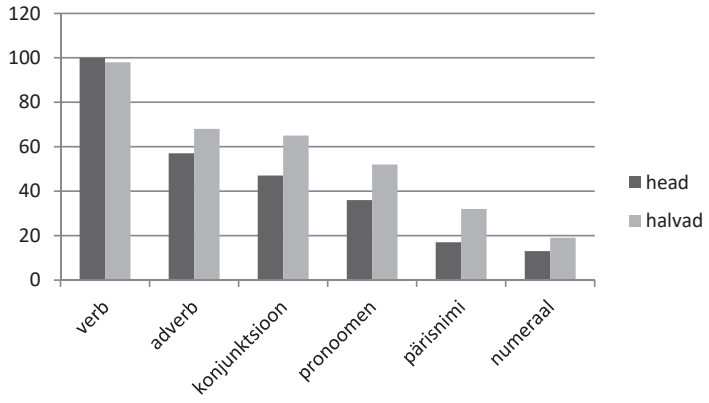
[11] Nimekiri pole lõplik ning sõnu võib tulevikus lisanduda.

Versioonis 1.2 on lause alguses keelatud adverbid *siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*, *siis*; versioonis 1.3 keelati need kogu lauses. See tähendab, et iga lause, kus esines vähemalt üks neist adverbidest, kaotas automaatselt 50% skoorist. Analüüsi käigus selgus, et adverbide *siin*, *siia*, *siit*, *seal*, *sinna* ja *sealt* keelamine lauses ei ole õigustatud (näited 42–45), küll aga tasub neid edaspidi taunida. Samas tõsteti pronoomeni esinemisele määratud karistust 0,1-lt 0,5-le. Adverb *siis* kustutati nimekirjast ja lisati lause alguses keelatud sõnade nimekirja (näide 46).

(42) Eesti riik on see, mis ühendab kõiki siin alaliselt elavaid inimesi.
(43) Isikukaardil on omaniku foto ja sõrmejälg, ka on seal kirjas päritolumaa.
(44) Jõin kummeliteed ja pigistasin sinna ka jõhvikaid, mis aitavad palaviku vastu.
(45) Ta ütles selgelt ja kindlalt – tema tahab siit varsti välismaale minna.
(46) Siis sadas seal lörtsi ja rahet.

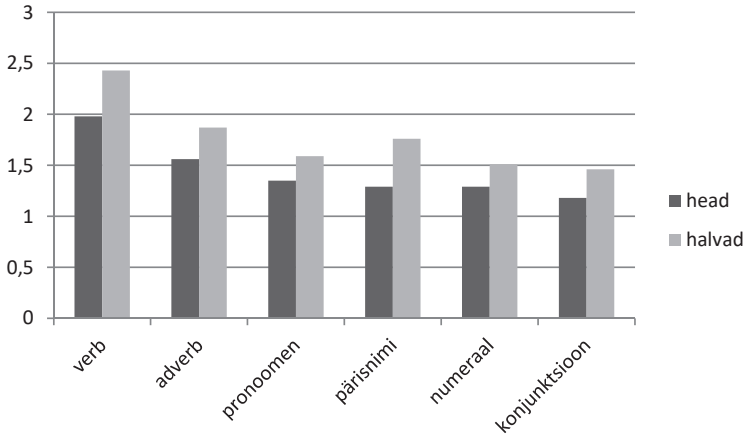
4.5. Lause liikmete sõnaliigiline kuuluvus ja esinemise arv lauses

Analüüsi käigus uuriti ka lause sõnaliigilist koosseisu. Esmalt selgitati andmebaaside analüüsi käigus välja, mitmel protsendil kõikidest lausetest esineb verb, adverb, pronoomen, pärisnimi, numeraal või konjunktsioon. Nii selgus nt, et 98% halbadest ja 100% headest näitelausetest sisaldavad verbi, 32% halbadest ja 17% headest näitelausetest sisaldavad pärisnime jne (vt joonis 8).



Joonis 8. Sõnaliigi esinemise protsent lausetes

Seejärel võeti arvesse ainult need laused, kus teatud sõnaliik (nt pronoomen) esineb, ning loeti kokku, mitu korda see kordub – nt kas lauses esineb ainult 1 pronoomen või 2, 3, 4, .. pronoomenit. Analüüsi tulemusena selgus nt, et halbades näitelausetes esineb keskmiselt 2,43 ja heades näitelausetes 1,89 verbi, halbades näitelausetes esineb keskmiselt 1,76 ja heades näitelausetes 1,29 pärisnime jne (vt joonis 9).



Joonis 9. Verbi, adverbi, pronoomeni, pärisnime, numeraali, konjunktsiooni keskmised esinemisjuhud lauses

Kuna õppesõnastike puhul on eelistatud lihtlaused, tehti kindlaks ka komade arv. Heade näitelausete andmebaasis on komadega lauseid 29%, halbade näitelausete andmebaasis 53%. Headel näitelausetel on vähem komasid: komadega lauses keskmiselt 1,13 koma, samas kui halbadel näitelausetel on 1,44.

Versiooni 1.4 jaoks lisati uued parameetrid nõrkade klassifikaatorite alla: taunitud on laused, kus on rohkem kui 2 verbi, 1 adverb, 1 pronoomen, 1 pärisnimi, 1 numeraal, 1 konjunktsioon, 1 koma.

Koppel ja Kallas (2016) tõdesid, et versiooni 1.3 väljundis olid pärisnimesid ja numeraale sisaldavad laused endiselt esil. Versiooni 1.4 lisati klassifikaator, mis vähendab iga pärisnime sisaldava lause skoori. See tähendab, et lause, kus esinevad pärisnimed, võib saada topelt karistada – juba ühe pärisnime esinemine lauses vähendab lause skoori (karistuseks on määratud 0,1). Kui pärisnimesid on lauses rohkem kui 1, saab lause veel täiendavalt karistada (karistus 0,2). Arvsõnade puhul kahte karistust ei rakendatud, sest vastasel juhul saaksid sellised laused nagu näited (47–48) madalama skoori:

(47) Üks kirp hammustab vere otsingul samas piirkonnas sageli kaks või kolm korda.
(48) Abielust sündis kolm last – tütar ja kaks poega.

4.6. Nõrkade klassifikaatorite kaal

Versioonides 1.2 ja 1.3 ei olnud nõrkadele klassifikaatoritele kaale määratud. Versioonis 1.4 keskenduti pärast iga klassifikaatori häälestuse testimist klassifikaatorite kaalule: millised nõrgad klassifikaatorid on olulisemad kui teised. Igale klassifikaatorile määrati oma kaal, v.a juhul, kui klassifikaatorid grupeeriti, kuna nad jagasid ühistest tunnustest tingituna sama kaalu (teatud elemendi, nt koma kordumine lauses). Pärast põhjalikku testimist ning tulemuste analüüsi osutusid olulisimaks lause pikkuse optimaalne vahemik ja sõnade sagedus.

Analüüsi tulemusena selgus, et see oli üks olulisemaid GDEX 1.4 väljundit mõjutanud muudatusi.

5. Tulemused ja edasiarendused

Tugevad GDEX 1.4 parameetrid:

- lause algab suure tähega ja lõpeb lauselõpumärgiga;
- lause pikkus on 4–20 sõnet;
- sõna maksimaalne pikkus on 20 tähemärki;
- lauses peab esinema verb;
- lauses ei esine keelatud kirjamärke;
- lause alguses on keelatud teatud sõnad, sõnapaarid, sõnaliigid ja suurtähtedega kirjutatud sõnad;
- lauses ei esine sõnavorme, mille sagedus korpuses on alla 5;
- märksõna ei kordu.

Nõrgad GDEX 1.4 parameetrid:

- taunitakse lauseid, kus esineb lemma, mille sagedus korpuses on alla 1000;
- lause pikkuse optimaalne vahemik on 6–12 sõnet;
- taunitakse lauseid, kus esinevad harvad kirjamärgid, pronoomenid, musta nimekirja kuuluvad sõnad, lühendid, pärisnimed, *mast-*, *mas-*, *maks-* ja *des*-lauselühendid, lemma *kroon* koos numeraaliga;
- taunitakse lauseid, kus esineb rohkem kui 2 verbi, rohkem kui 1 adverb, rohkem kui 1 pronoomen, rohkem kui 1 konjunktsioon, rohkem kui 1 pärisnimi, rohkem kui 1 numeraal ja rohkem kui 1 koma.

GDEX 1.4 konfiguratsioonifaili katkend on esitatud joonisel 10, kus on näha tugevad (vt read 2–14) ja nõrgad (vt read 16–32) klassifikaatorid koos neile määratud kaaludega.

```
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),
 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, bad_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, ..
 + 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
```

Joonis 10. Katkend GDEX 1.4 konfiguratsioonifailist

Võrreldes versiooniga 1.3 paranes versiooni 1.4 väljund märgatavalt (vt joonised 11 ja 12). Versiooni 1.4 väljundi eesotsas on lühemad laused (vt joonis 12).

Old rank	Rank	Sentence
1	43	Meelelahutustööstus oli veel sündimata ning kollast ajakirjandust asendas klatsh turuplatsil .
2	44	Erinevalt kollasest ajakirjandusest pole naisteajakirjad orienteeritud skandaalsete seikade esiletoomisele .
3	100	Nii et ei tuleks tahtmist lugeda elutarkust kollasest ajakirjandusest .
4	17	Äripäev võiks arvestada sellega ja vältima muutumist kollaseks ajakirjanduseks .
5	80	Miks üleüldse informeeritakse sellistest olukordadest kollast ajakirjandust , aga mitte liikluspolitseid ?
6	30	Ma ei hakka oma isiklikku elu afišeerima kollases ajakirjanduses .
7	57	Isegi kollane ajakirjandus kasutab seda teemat lugejaskonna köitmiseks harva .
8	77	Kas sa teed oma tööna uurivat või kollast ajakirjandust ?
9	33	Kollane ajakirjandus võib midagi üritada , ja kindlasti üritab Galojan ise .
10	52	Briti kvaliteetlehed on hakanud lugejate meelitamiseks kasutama järjest rohkem kurikuulsa kollase ajakirjanduse võtteid .

Joonis 11. Kollokatsiooni *kollane ajakirjandus* näitelaused GDEX 1.3 väljundis

Old rank	Rank	Sentence
39	1	Kogu lugu kõlbab ainult kollase ajakirjanduse veergudele .
90	2	Kollane ajakirjandus pidi paratamatult tekkima !
48	3	Ma ei võtaks kollase ajakirjanduse tähelepanu kiitusena .
14	4	Kollase ajakirjanduse peale pole õieti põhjust pahane olla .
73	5	Me elame suuresti kollase ajakirjanduse ajastul .
65	6	Vassiljevite perele kargas kallale kollane ajakirjandus .
68	7	Kollane ajakirjandus on Eesti demokraatia hinnaks .
47	8	Ma ei taha kommenteerida kollase ajakirjanduse pealkirju .
40	9	Inglismaal on kollane ajakirjandus ju teadagi milline .
20	10	Prostitutsioon Soomes on seni huvitanud peamiselt kollast ajakirjandust .

Joonis 12. Kollokatsiooni *kollane ajakirjandus* näitelaused GDEX 1.4 väljundis

Eesti mooduli arendamise järgmises etapis võiks testida klassifikaatorit, mis eelistab substantiiviga algavaid lauseid. Samuti on võimalik testida, kas väljund paraneb, kui jagada musta nimekirja kuuluvad sõnad veel eri gruppidesse, millest vulgarismid on lausest täielikult keelatud, kuid tundlikud sõnad vähendavad lause skoori.

Plaanis on täiendavalt testida parameetrit, mis eelistab lauseid, kus esineb kolmas kollokaat ehk kollokatsiooni kollokaat. See parameeter parandas oluliselt nii inglise (Kilgarriff jt 2008) kui sloveeni keele (Kosem jt 2013) GDEX-i väljundit. Kolmas kollokaat tähendab seda, et kõrgema skoori saavad laused, milles sisaldub veel üks kõrge esilduvusega kollokaat. Nt kollokatsiooni *raamatut lugema* puhul annab GDEX kõrgema skoori lausetele, milles esineb kõrge esilduvusega kollokaat *huvitav* (*huvitavat raamatut lugema*).

Testandmebaasid saab metaandmetele toetudes omakorda osadeks jagada, ehk luua iga sõnaliigi jaoks eraldi andmebaasid. Need andmebaasid annavad infot

eri sõnaliiki kuuluvate märksõnade käitumise kohta lauses. See info võimaldab GDEX-i versiooni häälestada vastavalt sõnaliigile. Nt kui analüüs näitab, et kui märksõnaks on verb ja see esineb tavaliselt lause esimeses pooles, siis saab GDEX-i konfiguratsioonifaili lisada märksõna asukohta arvestava klassifikaatori.

Tulevikus on plaanis eri versioonide evalveerimine ja selle tulemuste analüüs. Edaspidi rakendatakse GDEX 1.4 KOLS-i veebiliideses, kust kasutaja saab näitelauseid otse tekstikorpusest vaadata. GDEX 1.4 üks võimalikke rakendusi on luua eestikeelne keeleõpperakendus SkELL[12] (Baisa, Suchomel 2014), mis sisaldab GDEX-i abil välja valitud lauseid. Lisaks on kavas luua eesti mooduli eri versioonid eri keeleoskustasemetele, mis arvestaksid ka vastava taseme sõnavaraloendeid.

6. Kokkuvõte

Artiklis keskenduti tööriista Good Dictionary Example ehk GDEX eesti mooduli uusimale versioonile 1.4. GDEX 1.4 väljatöötamiseks analüüsiti testandmebaase, mis sisaldasid koostamisel oleva eesti keele kollokatsioonisõnastiku näitelauseid. Näitelausete parameetrite kvantitatiivsele analüüsile ning eri versioonide võrdlusele toetudes arendati välja GDEX-i eesti mooduli uus versioon 1.4.

GDEX 1.4 arendamisel testiti põhjalikult klassifikaatoreid, mida tööriist näitelausete tuvastamiseks kasutab. Tugevad klassifikaatorid sisaldavad parameetreid, millele lause peab vastama (nt lause peab olema täislause; lause peab sisaldama verbi). Nõrgad klassifikaatorid mõjutavad lause skoori vähem, neile mitte vastamise eest saab lause n-ö karistada (nt lause pikkuse optimaalne vahemik on 6–12 sõnet). Erinevalt varasematest versioonidest lisati GDEX 1.4 nõrkadele klassifikaatoritele oma kaalud. See osutus üks olulisemaks väljundit mõjutanud muudatuseks.

Eri versioonide võrdlemiseks ja parameetrite häälestamiseks kasutati spetsiaalset programmi nimega GDEX Editor. Testandmebaaside statistilise analüüsi tulemusi rakendades ja eri versioone võrreldes kustutati ebaolulisi, parandati vanu ning lisati juurde uusi parameetreid. GDEX-i väljund paranes märgatavalt.

GDEX 1.4 väljundit saab rakendada eesti kollokatsioonisõnastiku projektis ning SkELL-taolise keeleõpperakenduse loomisel eesti keele jaoks.

Viidatud kirjandus

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

Kaalep, Heiki-Jaan 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. [An Estonian morphological analyser and using a corpus on its development.] – Keel ja Kirjandus, 1, 22–29.

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

Kilgarriff, Adam; Husák, Milos; McAdam, Katy; Rundell, Michael; Rychlý, Pavel 2008. GDEX: Automatically finding good dictionary examples in a corpus. – E. Bernal, J. DeCesaris

[12] https://skell.sketchengine.co.uk (1.9.2016).

(Eds.), Proceedings of the 13th EURALEX International Congress. Barcelona: Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra, 425–432.
Kilgarriff, Adam; Rychlý, Pavel; Smr, Pavel; Tugwell, David 2004. The Sketch Engine. – G. Williams, S. Vessier (Eds.), Proceedings of the 11th EURALEX International Congress. Lorient, France: Université de Bretagne Sud, 105–115.
Koppel, Kristina; Kallas, Jelena 2016. Õppijasõbralik korpuslause: automaatse valiku võimalusi. [User-friendly corpus sentence: Parameters for automatic selection.] – Lähivõrdlusi. Lähivertailuja, 26, 222–250. http://dx.doi.org/10.5128/LV26.07
Kosem, Iztok; Gantar, Polona; Krek, Simon 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. Ljubljana–Tallinn: Trojina, Institute for Applied Slovene Studies, Eesti Keele Instituut, 32–48.
Schmid, Helmut 1994. Probabilistic part-of-speech tagging using decision trees. – Proceedings of International Conference on New Methods in Language Processing. Manchester, UK, 44–49.

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

AUTOMATIC DETECTION OF GOOD DICTIONARY EXAMPLES IN ESTONIAN LEARNER'S DICTIONARIES

Kristina Koppel
Institute of the Estonian Language, University of Tartu

This paper explains, firstly, how a tool called Good Dictionary Example (GDEX) (Kilgarriff et. al 2008) scores corpus sentences and helps the lexicographer automatically select the best examples for dictionaries. Secondly, the training datasets containing example sentences from the Estonian Collocations Dictionary (ECD) are introduced. Thirdly, the paper focuses on different parameters of good dictionary examples.

Most of the paper is based on an analysis of the training datasets and an evaluation of the previous GDEX configurations. For evaluating the configurations, the graphical user interface GDEX Editor was used. Based on the results of statistical analysis and on the evaluation of different configurations, a new configuration 1.4 is introduced. There are 16 new parameters implemented in GDEX 1.4.

The main parameters of GDEX 1.4 are as follows: the desired sentence is a full sentence; sentence length is 4–20 tokens; the sentence contains a verb; it does not contain low frequency words or words from the blacklist; the optimal length is 6–12 tokens; sentences containing more than 1 adverb, pronoun, proper name, numeral, conjunction, comma, more than 2 verbs and sentences containing certain pronouns are penalized.

The output of GDEX 1.4 can be applied to the ECD project and to create a web interface SkELL for learners of Estonian.

Keywords: corpus lexicography, corpus linguistics, learner's lexicography, language learning, collocations, usage examples, GDEX, Estonian