doi:10.5128/ERYa11.05

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 11, 75–94

KORPUSLEKSIKOGRAAFIA UUED VÕIMALUSED EESTI KEELE KOLLOKATSIOONISÕNASTIKU NÄITEL

Jelena Kallas, Kristina Koppel, Maria Tuulik

Ülevaade. Artiklis tutvustame korpusleksikograafia üldisi arengutendentse ja uusi meetodeid. Käsitleme korpuse kui leksikograafilise info allika potentsiaali ning analüüsime, kuidas saab leksikograafilisi andmebaase pool- ja täisautomaatselt genereerida. Vaatleme, mil määral on uusi tehnoloogilisi lahendusi võimalik rakendada Eesti õppeleksikograafias, täpsemalt eesti keele kollokatsioonisõnastiku (KOLS) koostamisel. KOLS on esimene eestikeelne sõnastik, kus rakendatakse andmebaasi automaatset genereerimist nii märksõnastiku kui ka sõnaartikli sisu (kollokatiivse info ja näitelausete) tasandil. Tutvustame sõnastiku koostamise üldisi põhimõtteid ja esitame näidisartikli.*

Võtmesõnad: korpusleksikograafia, kollokatsioonisõnastik, korpuspäringusüsteem, sõnastikusüsteem, eesti keel

1. Sõnastike korpuspõhine koostamine internetiajastul: vahendid ja meetodid

Korpusleksikograafia eesmärk on luua meetodid, mis võimaldaksid leksikograafilisi üksusi automaatselt tuvastada ja valida. Korpusanalüüsi tulemusi rakendatakse tänapäeval märksõnastiku loomisel, tähendusjaotuste uurimisel, leksikaalsete üksuste erinevate omaduste tuvastamisel (süntaktiline käitumine, kollokatsioonid, semantiline prosoodia eri tüüpi tekstides, leksikaal-semantilised suhted), näitelausete valikul ja tõlkevastete leidmisel (Kilgarriff 2013: 77).

Korpuspõhine koostamine eeldab suure hulga erineva sisuga korpuste olemasolu. Näiteks ei pruugi korpus, mis on piisavalt representatiivne õppesõnastiku koostamiseks, olla piisav suure seletava sõnaraamatu jaoks. Teiselt poolt on oluline, et korpuste sisu oleks võimalik pidevalt täiendada ja uuendada. Adam Kilgarriff (2001: 345) tõdes juba 2001. aastal, et korpuste loomiseks saab edukalt kasutada

* Eesti keele kollokatsioonisõnastiku koostamist toetab Haridus- ja Teadusministeeriumi riiklik programm "Eesti keel ja kultuurimälu II (2014–2018)". Sõnastik valmib 2018. aastal.

veebimaterjali. Veebi põhieelis on see, et tekstid on oma olemuselt autentsed ja juba masinloetavas formaadis. Geoffrey Leech (2007: 145) on osutanud, et korpus on esinduslik ainult siis, kui see on koostatud ja kujundatud hoolikalt valitud materjalist. Ilmselgelt ei saa seda öelda veebi kohta, mis koosneb laiaulatuslikest igapäevasituatsioone puudutavatest tekstidest ja mille sisu ei ole keeleteadlaste valitud. Maristella Gatto (2014: 43–45, 67) arvates see just veebi esinduslikuks teebki. Veeb pakub juurdepääsu mitmele žanrile, millest mõni, näiteks akadeemilised tekstid, on korralikud kirjakeelsed tekstid, kuid teised, näiteks blogid, on lähemad kõnekeelele.

Veebi dünaamilise iseloomu tõttu ei saa aga päringu kordamisel olla kindel, et tulemus on sama. See võib osutuda probleemiks veebi kasutamisel leksikograafilise allikmaterjalina. Üks võimalik lahendus on laadida otsingu tulemused alla ning luua neist stabiilne ja kontrollitav andmebaas (Gatto 2014: 68–69). Sellist lähenemist rakendatakse näiteks TenTen-korpuste sarja loomisel (Jakubíček jt 2013). TenTen osutab sellele, et ühes veebikorpuses võib olla kuni 10^{10} sõnet. TenTen-korpused on praegu olemas nii suurtele maailmakeeltele, nagu inglise, araabia, hiina, portugali, prantsuse, saksa, itaalia, jaapani, korea, vene ja hispaania keel, kui ka väiksematele, nagu tšehhi, ungari, poola, slovaki ja eesti keel. Eestikeelsete veebilehtede infost koosneva korpuse etTenTen13[1] koostas firma Lexical Computing Ltd. 2013. aasta sügisel. Korpuse on lemmatiseerinud, märgendanud ja ühestanud OÜ Filosoft. Ajakirjandustekstid moodustavad etTenTen13 korpusest 29%, foorumid ja blogid 23%, teabetekstid 9%, 4% moodustavad usulise ja poliitilise sisuga tekstid ning 35% on liigitamata tekstide osakaal[2].

TenTen-korpuste loomisprotsess on järgmine: esmalt otsitakse tarkvaraga SpiderLing (Pomikalek, Suchomel 2012) välja vastava keele veebilehed ja tõmmatakse tekstid üheks korpuseks kokku; seejärel kustutatakse mittetekstiline materjal ja korduvad tekstid programmidega JusText ja Onion (Pomikalek 2011). Sel viisil on korpuse loomine väga kiire – näiteks koguti 12 miljardit ingliskeelset sõna sisaldav enTenTen12 ainult 12 päevaga.

Kui korpus on loodud, installeeritakse see korpuspäringusüsteemi. Selliste süsteemide viimasesse põlvkonda kuulub näiteks Sketch Engine (Kilgarriff jt 2004; Kilgarriff, Kosem 2012), mida kasutatakse Eesti Keele Instituudis[3]. Programmi funktsioonid on konkordantsi koostamine ja selle mitmekülgne töötlemine, statistikapõhine kollokaatide leidmine, korpusest sõnaloendite ja sagedusloendite koostamine, sõna süntaktilist ja kollokatiivset käitumist illustreerivate sõnavisandite (Word Sketch) genereerimine, sõnastikunäidete automaatne valimine, tesauruse koostamine. Seega sisaldab tarkvara erinevaid funktsioone, mida saab konkreetse leksikograafilise projekti puhul rakendada. Osa funktsioonidest on universaalsed ja kergesti kohandatavad kõikidele keeltele (eeldusel, et korpus on lemmatiseeritud, morfoloogiliselt märgendatud ja ühestatud), osa funktsioone nõuab aga keelespetsiifiliste rakenduste loomist. Nii eeldab näiteks sõnavisandite genereerimine iga keele jaoks spetsiaalse sõnavisandite grammatika (Sketch Grammar) kirjutamist (eesti keele mooduli kohta vt lähemalt Kallas 2013: 31–87), heade näitelausete tuvastamiseks on samuti vajalik võtta arvesse mitmeid keele tüübist tingitud omadusi.

[1] Korpus on kättesaadav aadressil www.keeleveeb.ee ning programmi Sketch Engine https://the.sketchengine.co.uk/auth/corpora/ kaudu (29.9.2014). Korpuse nimetuses sisalduv number 13 osutab, et korpus loodi 2013. aastal.

[2] Vt lähemalt http://www2.keeleveeb.ee/dict/corpus/ettenten/about.html (29.9.2014).

[3] Programmi kasutati "Eesti keele põhisõnavara sõnastiku" (PSV) koostamisel, hetkel kasutatakse seda üheköitelise eesti keele seletava sõnaraamatu (Langemets jt 2010) ja eesti keele kollokatsioonisõnastiku koostamisel.

Korpuspõhisuse põhimõte tänapäeva leksikograafias on paratamatult avaldanud mõju kogu sõnastike koostamisprotsessile. Annette Klosa (2013: 520–522) eristab sealjuures kuut etappi: 1) projekti kavandamine (sh pilootuuring); 2) allikmaterjali määramine (tekstikorpused, ilmunud paber- või veebisõnastikud, grammatikaõpikud, leksikaalsed andmebaasid, pildipangad jmt); 3) sõnastiku andmebaasi struktuuri loomine; 4) allikmaterjali (eelkõige tekstikorpuste) installeerimine korpuspäringusüsteemi; 5) andmete töötlus; 6) veebiliidese loomine ja avalikustamine.

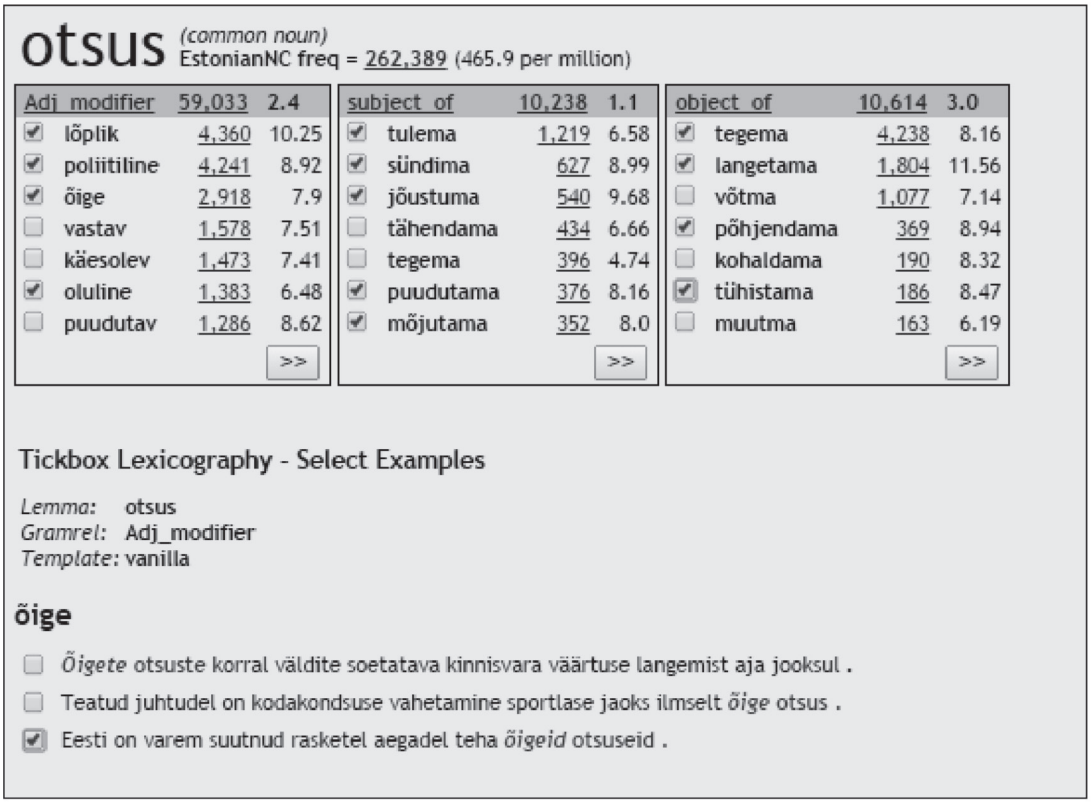
Juba enne sõnastiku loomist tuleb täpselt määrata, millisele korpusmaterjalile hakkab leksikograaf toetuma. Mõnikord tuleb vastav korpus alles luua, see märgendada ja installeerida korpuspäringusüsteemi. Koostamist saab oluliselt kiirendada, kui andmebaas genereerida pool- või täisautomaatselt. Seejuures võib tulemus olla kahesugune: a) info, mis enam toimetamist ei vaja, näiteks info märksõna sageduse kohta; b) info, mida tuleb käsitada toorandmestikuna ja mis vajab leksikograafilist järeltoimetamist. Sellised infoüksused on näiteks tähendusjaotused, definitsioonid, kollokatsioonid ja näitelaused.

Sõnastiku andmebaasi täisautomaatset genereerimist on rakendatud näiteks sloveeni keele leksikograafilise andmebaasi Slovene Lexical Database (Kosem jt 2013) ja inglise keele leksikaalse andmebaasi DANTE[4] genereerimisel. Meetod sobib eriti hästi kollokatsioonisõnastike andmebaaside koostamiseks. Automaatgenereerimine toetub tarkvara Sketch Engine sõnaloendi (Word List), sõnavisandi (Word Sketch) ja heade näitelausete (Good Dictionary Example ehk GDEX) funktsioonidele. Sõnaloend võimaldab sageduspõhise märksõnastiku loomist, päringut saab teha nt sõnavormipõhiselt, sõnaliigipõhiselt või grammatiliste tunnuste järgi. Sõnavisandi abil saab tuvastada sagedamaid ja kõrge esilduvusega kollokatsioone. GDEX võimaldab leida sobivad näitelaused. Kosemi jt (2013: 41–42) uurimus näitas, et andmebaasi täisautomaatne koostamine vähendab leksikograafi ajakulu umbes poole võrra.

Poolautomaatset koostamist võimaldab Sketch Engine'i Tickbox Lexicography (TBL) meetod (Kilgarriff jt 2010), mis seisneb selles, et leksikograaf valib sõnavisandit analüüsides konkreetse lekseemi jaoks sobivad kollokaadid ja näitelaused ning seejärel toimub nende automaatne ülekanne sõnastikusüsteemi. Nii on koostatud näiteks "Macmillan Collocations Dictionary for Learners of English" (Rundell 2012). Meetodit rakendatakse ka suure hollandi keele sõnaraamatu "Algemeen Nederlands Woordenboek" (Tiberius, Schoonheim (ilmumas)) koostamisel.

Joonis 1 illustreerib TBL-i võimalikku kasutust eesti substantiivi *otsus* näitel. Leksikograaf märgib linnukesega, mis kollokaadid sõnaartiklisse sobivad, siis valib iga kollokatsiooni jaoks sobivad laused (joonisel on näidatud kollokatsiooni *õige otsus* näitelaused), seejärel salvestab tulemuse ja kopeerib valitud üksused sõnastikusüsteemi.

[4] Vt http://www.webdante.com/ (29.9.2014).



Joonis 1. Kollokaatide ja näitelausete valik Tickbox Lexicography meetodil

Poolautomaatse koostamise eelis on see, et märksõnu saab valida n-ö jooksvalt, sõnastiku koostamise käigus. Teisalt võimaldab see meetod väljundit kontrollida ja paremini jälgida terminite, harva esinevate sõnade, slängi ja erisuguse müra sattumist kollokaatide hulka ja näidetesse.

Siinses artiklis vaatleme, mil määral saab automaatset genereerimist rakendada Eesti õppeleksikograafias, täpsemalt, eesti keele kollokatsioonisõnastiku koostamisel. Selle andmebaasi automaatse genereerimise alus on 2014. aastal loodud eesti keele ühendkorpus (Estonian National Corpus)[5], mille struktuuri tutvustame lähemalt peatükis 3.

2. Eesti keele kollokatsioonisõnastiku koostamispõhimõtted

Eesti keele kollokatsioonisõnastik (KOLS) on Eesti Keele Instituudi 2014. aastal alanud projekt, mille sihtgrupp on peaasjalikult eesti keele õppijad (B2-C1 keeleoskustase), kuid ka emakeelsed kõnelejad.

Igal keelel on eriomaseid sõnade kombinatsioone, mille tundmine on vajalik, et oleks võimalik selles keeles loomulikult ja ladusalt rääkida ning kirjutada. Näiteks on eesti keeles täiesti normaalne öelda *tugev vihm*, kuid inglise keeles see nii ei ole (öeldakse *heavy rain*, aga mitte *strong rain*). Kollokatsioonisõnastike eesmärk ongi aidata keeleõppijal valida õigeid keelendeid, mis väljendaksid nende mõtteid loomulikul moel ja teeksid nende teksti sarnaseks emakeelse kõneleja omale.

[5] Vt https://the.sketchengine.co.uk/auth/corpora/ (29.9.2014). Praegu on ühendkorpus kättesaadav ainult korpuspäringu tarkvara Sketch Engine kaudu.

Batia Laufer (2011: 44) tõdeb, et paljud keeleõppijad ei ole kollokatsioonidest teadlikud – nad ei tea, et näiteks teatud substantiivid esinevad vaid koos piiratud arvu verbidega. Õppijad võivad aga valida verbi, mis nende emakeeles konteksti sobitub, kuid sihtkeeles on vale. Üsna tihti ei leia kasutajad infot kollokatsioonide kohta isegi siis, kui see on sõnastikus olemas (nt näitelausete tasandil). Robert Lew (2004: 23) märgib, et enamik algajaid õppijaid ei otsigi sõnastikest kollokatsiooniinfot, vaid see huvitab rohkem edasijõudnuid, kellel on suurem teadmine keelestruktuurist.

Meie käsitleme kollokatsioonidena sisusõnade tähenduslikke ja statistiliselt esilduvaid kombinatsioone teiste leksikaalsete ja grammatiliste üksustega. Kollokatsioonid on nt *ere päike*, *päike paistab*, *päikest võtma*.

Eesti keele kollokatsioonisõnastiku põhilised infoüksused on märksõna, definitsioon (vaid mitmetähenduslike sõnade puhul), kollokatsioonid ja näitelaused. Tähendusi eristame ainult siis, kui see on kollokatsioonide selgema esituse huvides vajalik. Põhjalik polüseemia kirjeldamine pole kollokatsioonisõnastikus eesmärk omaette.

KOLS-i maht on umbes 10 000 märksõna. Võrdluseks olgu siinkohal toodud "Macmillan Collocations Dictionary for Learners of English" (MCD 2010), kus märksõnu on 4500 ja kollokatsioone rohkem kui 121 000, ning "Oxford Collocations Dictionary for Students of English" (OCDSE 2002), kus märksõnade hulk on 9000, kollokatsioone kokku 150 000 ja näitelauseid rohkem kui 50 000.

Reeglina on kollokatsioonisõnastike märksõnadeks vaid sisusõnad. Substantiivid, adjektiivid ja verbid moodustavad kuni 99% ja määrsõnad vähem kui 1% kogu märksõnastikust (Kilgarriff jt 2014).

KOLS-i märksõnastikku kuuluvad substantiivid, adjektiivid, verbid ja lisaks valikuliselt ka viisiadverbid (nt *salaja*, *kiiresti*). Eraldi kontrollime, et märksõnastikus on kindlasti ka akadeemilises kirjutamises vajaminev sõnavara (Metslang, Kibar 2012)[6]. Mitmesõnaliste märksõnadena esitame semantiliselt terviklikke, omaette tähenduse ja argumentstruktuuriga ühendverbe (*kaasa lööma*, *alla andma*), väljendverbe (*silma paistma*, *silmas pidama*) ja isolaadi ehk ainukordse komponendi ja verbi ühendeid (*andeks andma*, *tähele panema*, *kihla vedama*, *toime tulema*).

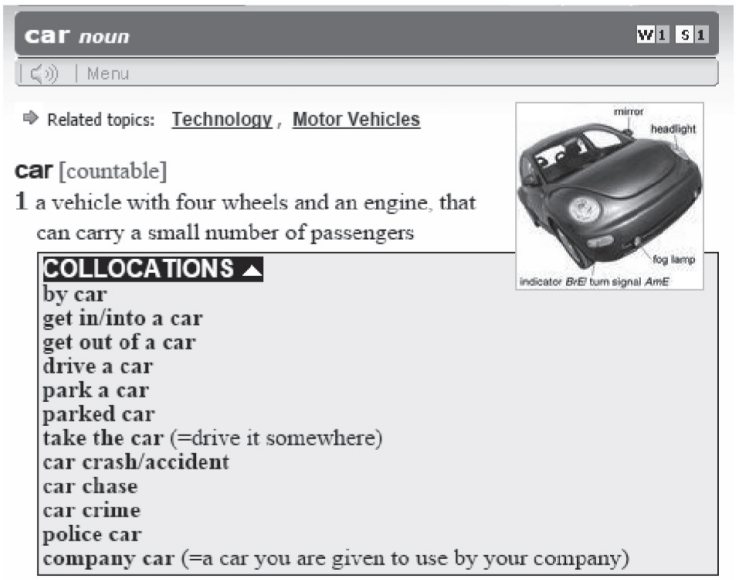
Tänapäeva inglise sõnastikes (nt OCDSE 2002, MCD 2010, "Longman English Dictionary"[7], "Collins English Dictionary"[8]) esitatakse kollokatsioone mitmel viisil:

- kollokatsioonirühm on esitatud märksõna all sõnaliigipõhise koodi ja/või sümboli taga. Rühmitamise aluseks on kollokatsioonikood, nt Adj+N (adjektiiv + substantiiv);
- kollokatsioonid esitatakse loendina eraldi kastikeses (vt joonis 2);
- kollokatsioonid tuuakse näidete sees esile teistsuguse fondi abil;
- kollokatsioonid esitatakse näidete sees eraldi esile toomata;
- kollokatsioonid esitatakse definitsiooni osana.

[6] 400-sõnalise üldakadeemilise sõnavara loend on koostatud ülikoolis õpetatavate erialade tekstide põhjal. Loend ei ole valdkonnaspetsiifiline, vaid sellesse on valitud üldisemad sõnad, mis on paljudel erialadel ühised.

[7] Vt http://www.ldoceonline.com/ (29.9.2014).

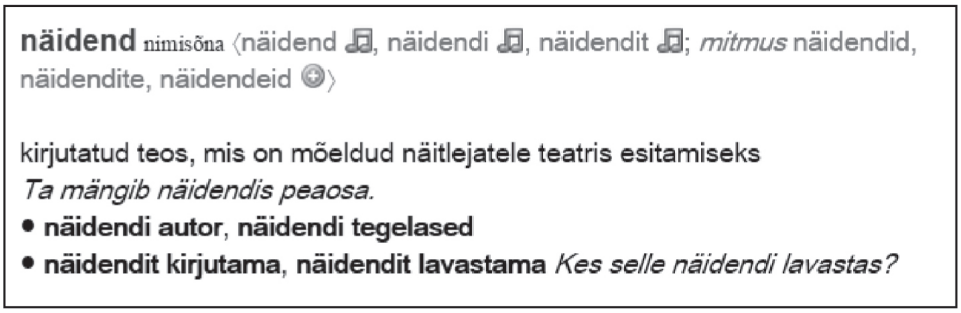
[8] Vt http://www.collinsdictionary.com/dictionary/english-cobuild-learners (29.9.2014).



Joonis 2. Kollokatsioonide esitus sõnastikus "Longman English Dictionary"

Süntagmaatilise info esitamise varasemad uurimused eesti keele üld- ja õppesõnaraamatutes (Langemets jt 2005, Kallas, Tuulik 2011, Kallas 2013) on näidanud, et eesti leksikograafias ei ole välja kujunenud kollokatsioonisõnaraamatute koostamise traditsiooni. Kollokatsioone pole tavaks esitada eksplitsiitselt, enamasti tehakse seda näitelausete tasandil.

Üks esimesi katseid kollokatsioone eksplitsiitselt esitada tehti "Eesti keele põhisõnavara sõnastikus" (PSV). Kasutaja jaoks eristati kollokatsioonirühmad muust artiklist mummu abil ja paksu kirjaga (vt joonis 3), andmebaasis aga koodidena (vt lähemalt Kallas 2013: 104–126).



Joonis 3. Kollokatsioonide esitus PSV-s

KOLS-i koostamisel järgime PSV tarbeks välja töötatud kollokatsioonide esituspõhimõtteid, üritame hoiduda liigsest lahterdamisest, esitame koos läbipaistvate kollokatsioonidega (nt *päikest nautima*) ka läbipaistmatumaid (nt *päikest võtma*) ja lisame vajadusel seletuse.

Kollokatsioone rühmitame sõnaliikide ning morfoloogiliste ja süntaktiliste kategooriate põhjal järgmistesse kollokatsioonirühmadesse (vt tabel 1).

Tabel 1. KOLS-i kollokatsioonirühmad (koos näidetega)

Substantiivi mallid	
adjektiiv + substantiiv	*hea/ilus/kõlav laul*
substantiiv (genitiivis) + substantiiv	*ekspertide hinnang* *koosoleku otsus*
substantiiv + substantiiv (partitiivis)	*viil leiba/juustu/saia*
substantiiv (adverbiaalkäändes) + substantiiv	*kullast ehted/kõrvarõngad* *osavõtt konkursist/võistlustest/valimistest*
substantiiv (subjekti funktsioonis) + verb	*hobune hirnub* *palavik tõuseb/langeb*
substantiiv (objekti funktsioonis) + verb	*arvutit sisse lülitama/välja lülitama*
substantiiv (adverbiaali funktsioonis) + verb	*aktsiatesse investeerima*
substantiiv + adpositsioon(ifraas)	*lepingu kohaselt/järgi* *uhkus saavutuste/tehtu üle*
adverb + substantiiv	*raagus puud* *omaette tuba*
substantiiv + verb *ma*- või *da*-infinitiivis	*meister valetama* *soov laulda*
rinnastustarind võrdlustarind	*päike ja tuul* *elu kui kabaree*
Adjektiivi mallid	
adjektiiv + substantiiv	*raske otsus* *rõõmsates toonides/värvides* *rõõmsal häälel*
adverb + adjektiiv	*väga aeglane* *silmatorkavalt hea*
adjektiiv (translatiivis) + verb adjektiiv (essiivis) + verb adjektiiv (nominatiivis) + verb	*rikkaks saama* *kahtlasena paistma* *ilus välja nägema*
adjektiiv + verb *ma*- või *da*-infinitiivis	*raske mõista* *pädev otsustama/hindama*
adjektiiv + adjektiiv	*igavene suur*
adjektiiv + adpositsioonifraas	*kingituste üle rõõmus* *hull raha järele*
rinnastustarind võrdlustarind	*rikas ja ilus* *valge kui lumi* *must nagu süsi*
Adverbi mallid	
adverb + adverb	*aina rohkem* *väga kiiresti*
adverb + adjektiiv	*väga aeglane*
adverb + verb	*kiiresti jooksma*
adverb + substantiiv	*palju rahvast*
adverb + adpositsioonifraas	*selja pealt lõhki* *puusade ümbert pingul*
rinnastustarind võrdlustarind	*hästi ja kiiresti* *kergelt kui õhk*

Verbi mallid	
adverb + verb	*kiiresti jooksma*
substantiiv (subjekti funktsioonis) + verb	*päike tõuseb/loojub*
substantiiv (objekti funktsioonis) + verb	*lilli istutama/kastma/korjama/kinkima*
substantiiv (adverbiaali funktsioonis) + verb	*lugupidamisega/austusega/eelarvamusega suhtuma*
adjektiiv (translatiivis) + verb adjektiiv (essiivis) + verb adjektiiv (nominatiivis) + verb	*paksuks minema* *võimatuna tunduma* *kummaline näima*
finiitverb + infiniitverb	*ajab naerma/nutma/iiveldama/oksendama* *jätab maksmata/tegemata*
verb + adpositsioonifraas	*vapruse/ülbuse/jutukuse poolest silma paistma*
rinnastustarind võrdlustarind	*kirjutama ja lugema* *elada või surra*

Andmebaasis jagame kollokatsioonid esitatud rühmade kaupa, kuid sõnastiku metakeeleks on vaid kollokaatide sõnaliigitähised välja kirjutatud kujul. Kollokatsioonide puhul arvestame kollokaatide vormide sagedust ja esitame neid kõige tüüpilisemal kujul, mitte tingimata algvormis. Seega omandab õppija koos kollokatsiooniga ka grammatilise info viisil, mida on lihtne kohe kasutusse võtta. Järgnevalt illustreerime kavandatavat kollokatsioonide esitust substantiivi *arutlus* näitel (1).

(1) **ARUTLUS** nimisõna

OMADUSSÕNAD

- **teoreetiline, avalik, pikk, filosoofiline, loogiline, tõsine, huvitav, sisuline, põhjalik** arutlus

NIMISÕNAD

- arutluse **objekt, tulemus, taust, tase**

TEGUSÕNAD

- arutlus **käib, toimub, algab, keskendub** *millele*, **jätkub, kestab, tekib**
- arutlust **jätkama, alustama, korraldama, kuulama, juhtima**
- arutlusele **tulema, võtma**
- arutlusel **olema**

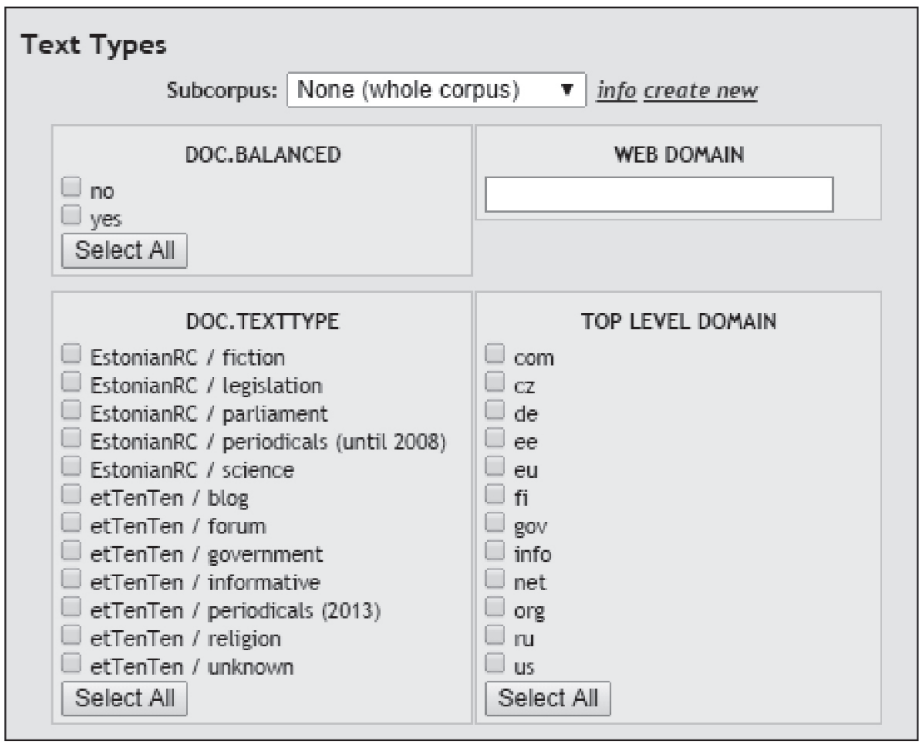
KAASSÕNAD

- arutluse **alla** (tulema, võtma)
- arutluse **all** (olema)

Plaanis on esitada iga kollokatsiooni juures näitelauseid, mis avaneksid kollokaadile klõpsates. Enamasti kasutame autentseid korpusnäiteid või nende mugandatud versioone.

3. KOLS-i andmebaasi allikas: eesti keele ühendkorpus

Eesti keele ühendkorpuse suurus on *ca* 563 mln sõnet ning hetkel on see suurim ja žanriliselt mitmekesiseim eesti keele korpus. Korpus koosneb eesti keele koondkorpusest (250 mln sõnet), sh tasakaalus korpusest (15 mln sõnet), ja eesti veebikorpusest etTenTen13 (*ca* 330 mln sõnet). Tarkvara Sketch Engine abil saab korpuse sisu analüüsida kas kogu korpuse ulatuses, allkorpuste või domeenide kaupa (vt joonis 4).



Joonis 4. Eesti keele ühendkorpuse allkorpused ja domeenid

Koondkorpuse allkorpused on ajakirjandustekstid kuni aastani 2008 (*periodicals* (*until 2008*)), ilukirjandustekstid (*fiction*), riigikogu stenogrammid (*parliament*), Eesti ja Euroopa seadused (*legislation*) ja teadustekstid (*science*). Veebikorpuse etTenTen13 allkorpused on ajakirjandustekstid (*periodicals* (2013)), foorumid (*forum*), blogid (*blog*), teabetekstid (*informative*), usutekstid (*religion*), ametlikud tekstid (*government*) ja varia (*unknown*). Leksikograafi vaatevinklist aga täiendavad koondkorpus ja veebikorpus teineteist: kui koondkorpuse tekstivalik võimaldab analüüsida eelkõige kirjakeelt, siis veebikorpus annab parema ülevaate just kõnekeele ja internetikeele kohta.

KOLS-i andmebaasi aluseks on ühendkorpus tervikuna, kuid soovi korral on toimetamise faasis võimalik uurida kollokatsioonide esinemist allkorpuste kaupa.

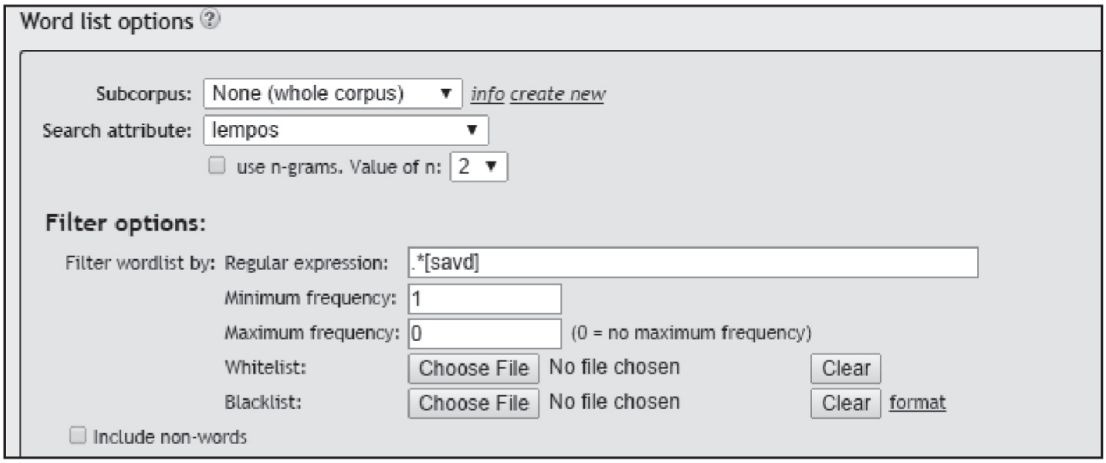
4. KOLS-i andmebaasi korpuspõhine automaatne genereerimine

KOLS-i andmebaasi genereerimine toimus korpuspäringutarkvaraga Sketch Engine ja koosnes järgmistest etappidest: 1) märksõnaloendi genereerimine sõnaloendi (Word List) funktsiooni abil, millele järgnes käsitsi kontroll; 2) sõnavisandite grammatika (Sketch Grammar) täiendamine ja andmete ekstraheerimiseks vajalike parameetrite täpsustamine; 3) näitelausete klassifikaatorite väljatöötamine ja GDEX-i skripti kirjutamine[9]; 4) spetsiaalse rakendusprogrammi kirjutamine ja andmete ekstraheerimine korpuspäringuprogrammist XML-faili kujul; 5) andmete importimine Eesti Keele Instituudi sõnastikusüsteemi EELex[10].

[9] Skripti süntaksi kohta vt https://www.sketchengine.co.uk/documentation/wiki/GDEX/Syntax (5.1.2015).
[10] Vt http://eelex.dyn.eki.ee/ (29.9.2014).

4.1. Märksõnastik

KOLS-i märksõnastiku koostamisel kasutasime Sketch Engine'i sõnaloendi (Word List) funktsiooni, mis võimaldab regulaaravaldiste abil erisuguse sisuga loendite genereerimist. Joonis 5 näitab funktsiooni päringuakent. Sealt nähtub, et loendit koostades otsib programm läbi kogu korpuse, otsinguatribuut on *lempos* (lemma + sõnaliik), otsitakse vaid substantiiviks, adjektiiviks, verbiks ja adverbiks märgendatud sõnu, lemma minimaalne ettemääratud esinemissagedus korpuses on 1, maksimaalne sagedus ei ole piiratud.



Joonis 5. Programmi Sketch Engine sõnaloendite funktsiooni päringuaken

Märksõnastiku aluseks võtsime 12 000 sagedasemat sõna, mida käsitsi kontrollisime. Kontroll oli vajalik näiteks selleks, et korpuse märgendamise kvaliteedist ja puudulikust ühestamisest tulenevat "müra" kõrvaldada. Tuli eemaldada vale variant märksõnadest, millel esines kaks kirjapilti (*mänedžer* vs. *mänedzher*, *šokk* vs. *shokk*, *režiim* vs. *rezhiim*), lühendid (*eek*, *eur*, *toim*), pärisnimed (*lõunaleht*, *suurhall*), liitsõna järelosised (*sugune*, *keelne*), vigased liitsõnad (*lapseoode*, *minumeel*, *omavahend*) ja terminid (*süsinikdioksiid*). Lisasime ka "Eesti keele põhisõnavara sõnastiku" (PSV) märksõnad, mis sagedusloendisse ei sattunud (*pott*, *köhima*, *hakkliha*).

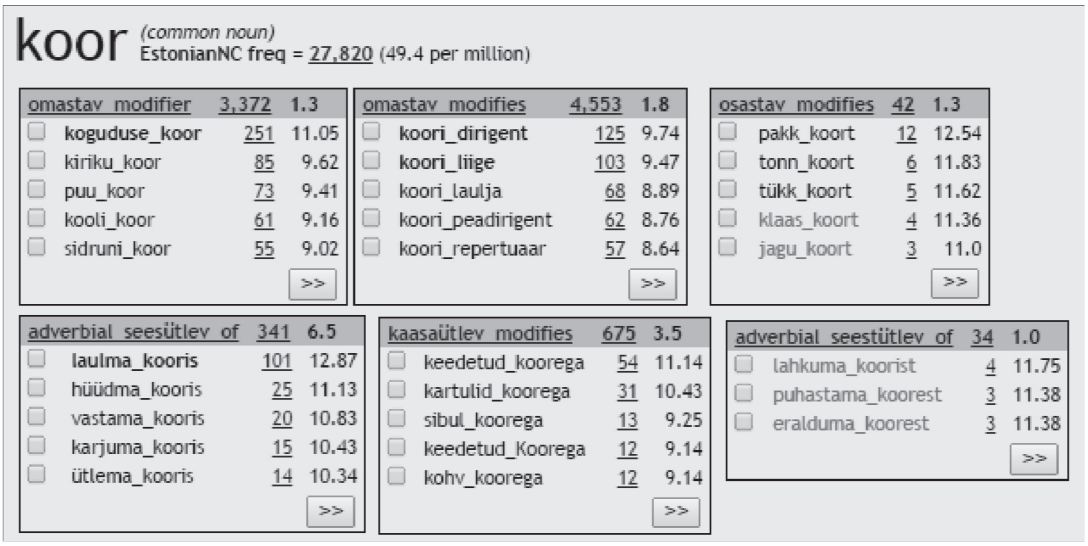
Märksõnade loendi automaatgenereerimise üks puudus on see, et pole võimalik tuvastada mitmesõnalisi märksõnu. Nende lisamine toimub käsitsi sõnastiku koostamise käigus. Kasutame kandidaatide valikul EKSS-i ja PSV perifrastiliste verbide loendit ning analüüsime sõnavisandite abil tuvastatud väljend- ja ühendverbe. Sõnavisandite grammatika võimaldab tuvastada verbi ja X-iks märgendatud sõnade koosesinemisi (nt *pärit olema*, *tähele panema*, *andeks andma*) ning verbi ja afiksaaladverbi funktsioonis esinevate sõnade koosesinemisi (nt *kaasa lööma*, *läbi lööma*, *välja lööma*, *kokku lööma*, *lahku lööma*). Samuti toetume eesti keele verbikesksete püsiühendite andmebaasile[11] ja tasakaalus korpuse lemmade ja sõnavormide mitmikute (n-grammide) sagedusloenditele[12].

[11] Vt http://www.cl.ut.ee/ressursid/pysiyhendid/index.php?lang=et (21.5.2014).
[12] Vt http://www.cl.ut.ee/ressursid/mitmikud/ (21.5.2014).

4.2. Kollokatiivne info

KOLS-i kollokatiivse info allikas on Sketch Engine'i sõnavisandi (Word Sketch) funktsiooni väljund (Kilgarriff jt 2004; eesti mooduli kohta vt lähemalt Kallas 2013: 31–87). Sõnavisand on spetsiaalse sõnavisandite grammatika abil genereeritud kokkuvõte sõna süntaktilisest ja kollokatiivsest käitumisest. 2013. aastal valminud eesti keele sõnavisandite grammatika versioon 1.5 sisaldas 85 reeglit ning selle väljund oli lemmapõhine, see tähendab, et kõik tuvastatavad kollokaadid olid sõnavisandis viidud lemmadele.

2014. aastal loodud grammatika versioonis 1.6 suudab programm tuvastada kollokaate nii lemma- kui ka sõnavormipõhiselt. On lisatud reeglid, mis võimaldavad tuvastada adverbiaali funktsioonis esinevate substantiivide ja verbide kollokatsioone (nt *konkursil/võistlusel/valimistel/võistlustel osalema*), käändumatuteks adjektiivideks märgendatud sõnade ja substantiivide kollokatsioone (nt *vastamata küsimus/kõne*) ning verbi ja nominatiivis adjektiivide kollokatsioone (*uskumatu/imelik/kummaline/võimatu tunduma*). Uues versioonis on kokku 116 reeglit, mille hulgas on 16 *unary*-tüüpi reeglit (võimaldavad analüüsida substantiivide ja adjektiivide morfoloogiliste vormide kasutussagedust), 4 *symmetric*-tüüpi reeglit (tuvastavad rinnastus- ja võrdlustarindeid, nt *päike ja tuul*, *ilus ja noor*, *valge nagu lumi*), 16 *dual*-tüüpi reegit (võimaldavad otsida kahe lemma koosesinemisi, nt *päike + paistma/loojuma/tõusma*) ja 80 *colloc*-tüüpi reeglit (võimaldavad tuvastada esiteks kolmest sõnast koosnevaid kollokatsioone, nt *uhke tehtu üle*, *hoolitsema laste eest*, ja teiseks kollokatsioone, kus otsisõna ja/või kollokaat ei esine lemmana, nt *kari lambaid/hobuseid/lehmi*, *rääkima aktsendita*, *suhtuma austusega/lugupidamisega* jne). Eriti kasulik on sõnavormipõhine esitus homonüümide puhul. Joonis 6 illustreerib valikuliselt substantiivi *koor* kollokatsioone.



Joonis 6. Valik substantiivi *koor* kollokatsioone

Jooniselt 6 on näha, et tuvastatud kollokatsioonid on nt *kooris laulma*, *kooris vastama/ütlema*, *koorist lahkuma*, *kooriga liituma*, aga *koorega kartulid/sibul*, *koorest puhastama/eralduma* jne.

Sõnavisandite grammatika versioon 1.6 võimaldab tuvastada kõiki KOLS-is esitatavaid kollokatsioonitüüpe (vt tabel 1).

4.3. Näitelaused

KOLS-i näitelausete alus on Sketch Engine'i funktsiooni GDEX (Kilgarriff jt 2008) väljund. GDEX on tööriist, mis hindab lausete kvaliteeti ja aitab leksikograafil leida korpusest parimad laused (Kilgarriff jt 2008: 425). Funktsioon töötab just-kui sõelana, hinnates lause süntaktilisi ja leksikaalseid tunnuseid ning sortides konkordantse selle järgi, kui hästi need hea lause kriteeriumitele vastavad. Tulemusena pakub tööriist nimekirja näitelausetest, kus eesotsas on head ja lõpupoole halvemad kandidaadid.

Esimeses GDEX-i versioonis inglise keele jaoks kasutati lausete kvaliteedi mõõtmiseks järgmisi klassifikaatoreid (Kilgarriff jt 2008: 426–427):

- lause pikkus on 10–25 sõna;
- lauses esinevad ainult sõnad, mis kuuluvad 17 000 sagedasema sõna hulka;
- lauses ei esine pronoomenid või anafoorid, nagu *this*, *that*, *it* või *one*;
- lause algab suure tähega ja lõppeb kirjavahemärgiga;
- kollokatsioon esineb lause lõpus. Eeldati, et hea näitelause tutvustab esmalt konteksti ja alles siis esineb kollokatsioon, mis sellesse konteksti sobitub. Nii saab kasutaja kollokatsiooni tähenduse konteksti põhjal tuletada, kui ta seda ei tea.

Eelpool nimetatud parameetreid sloveeni keele peal testides selgus, et GDEX-i poolt valitud lausete kvaliteet ei olnud piisavalt hea (Kosem jt 2011: 154–156, Kosem jt 2013: 38–39). Uurimistulemused näitasid, et GDEX nõuab iga keele jaoks eraldi konfiguratsiooni. Nii peeti sloveeni keele puhul kõige olulisemateks parameetriteks lause pikkust (8–30 sõna), väikese sagedusega sõnade läve (minimaalne sagedus korpuses ei tohi olla väiksem kui 3) ning pärisnimede ja pronoomenite puudumist lausetes. Samuti oli oluline märksõna asukoht. Kui märksõna asus lause alguses, ei olnud lause piisavalt informatiivne. Leiti, et on otstarbekas automaatselt eemaldada laused, kus esinevad lemma kordus ja erinevad sümbolid (meiliaadressid, püsilingid). Kvaliteedi parandamiseks oli koostatud ka eraldi loend sõnadest, mis ei tohtinud lauses esineda (nn must nimekiri), näiteks släng ja ropud sõnad. Lisaks määrati Levenshteini distants[13], mis tagas, et laused erinesid vähemalt 30% ulatuses. Hilisem uurimus tõi esile veel ühe tõsiasja: GDEX-iga saab paremaid tulemusi, kui kavandada iga sõnaliigi jaoks eraldi konfiguratsioon.

Selgitamaks eesti keele GDEX moodulile sobivaid parameetreid, võrdlesime "Eesti keele põhisõnavara sõnastiku" (PSV) ja koostamisel oleva üheköitelise eesti keele seletava sõnaraamatu[14] (SS1) näitelauseid. Võrdlusmaterjalina kasutasime etTenTen13 korpuse lauseid. Mõõtsime märksõnade (substantiivide, adjektiivide, adverbide ja verbide) parameetreid, nagu sõnade arv lauses, lausete keskmine pikkus, sõna keskmine pikkus ja kõrvallausetega lausete arv. Iga sõnaliigi puhul vaatasime 50 näitelauset. Seega analüüsisime kokku 600 lauset.

PSV laused on leksikograafide koostatud didaktilised üksused, mille eesmärk on näidata sõna kasutust kontekstis. Üheköitelise eesti keele seletava sõnaraamatu

[13] IT-s ja arvutiteaduses tähistab Levenshteini distants valemit, mille abil arvutatakse kahe järjendi vahelist erinevust.
[14] Elektrooline käsikiri Eesti Keele Instituudi sõnastikusüteemis EELex. Sõnastik valmib 2018. aastal.

(Langemets jt 2010) sihtrühm ei ole keeleõppijad, vaid emakeelsed haritud kasutajad. Sõnastiku näitelaused on suures osas adapteeritud (lauseid on lühendatud, välja on jäänud pärisnimed jmt). Veebikorpuse laused on autentne keelematerjal, mida ei ole leksikograafiliselt töödeldud. Tabelid 2 ja 3 võtavad kokku analüüsi tulemused.

Tabel 2. PSV ja SS1 näitelausete ning etTenTen13 korpuse lausete parameetrid

	Sõnade arv lauses	**Keskmine lause pikkus sõnades**	**Keskmine sõna pikkus (tm)**
Substantiivid			
PSV	3–9	5,08	5,6
SS1	3–12	6,42	6,7
etTenTen13	**4–40**	**15,8**	5,2
Adjektiivid			
PSV	3–10	5,08	5,3
SS1	5–11	6,44	6,7
etTenTen13	**3–37**	**15**	5,23
Verbid			
PSV	3–7	4,36	6,21
SS1	2–10	4,72	5,66
etTenTen13	**6–56**	**16,9**	6
Adverbid			
PSV	3–11	5,44	4,96
SS1	3–13	5,74	6,1
etTenTen13	**7–42**	**16,8**	5,64

Parameetrite kvantitatiivne analüüs toob selgelt esile lausete eripära. Õppeotstarbelise PSV näitelaused on tavaliselt üsna lühikesed (sõnade maksimumhulk on kuni 11, lauses on 4,36–5,44 sõna). SS1 laused on samuti üsna lakoonilised: sõnade maksimumhulk ulatub 13-ni ja lause pikkus on 4,72–6,44 sõna. Hoopis teised parameetrid on autentsetel, korpusest pärit lausetel ja vahe on väga suur: sõnade hulk lauses ulatub 56-ni ja lause keskmine pikkus on 15–16,9 sõna.

Näited (2–4) illustreerivad lauseid nimetatud allikates:

(2) Peremees kütab ahju. (PSV)
(3) Ta on mees parimais aastais. (SS1)
(4) Ükski ema ei anna vabatahtlikult oma last ära, selleks peab olema väga mõjuv põhjus, haigus, võimetus toime tulla või mida iganes. (etTenTen13)

Sõna keskmises pikkuses erilist vahet ei ole: see varieerub 4,96 tähemärgist 6,7 tähemärgini. Samas võivad eestikeelsed sõnad olla üsna pikad, nt *kirjanduslikustatamatumatelegi* (30 tähemärki), nii et ilmselt on otstarbekas sõnade maksimaalse pikkuse seadistamine.

Kõrvallausete analüüs näitas, et osalausetega lausete osakaal on üsna väike PSV-s ja SS1-s, samas autentses etTenTen13 korpuses oli osalausetega lausete osakaal tunduvalt suurem (substantiivide puhul 18%, adjektiivi puhul 58%, verbide ja adverbide puhul aga koguni 76%) (vt tabel 3).

Tabel 3. Kõrvallausetega lausete osakaal PSV-s, SS1-s ja korpuses etTenTen13

	Kõrvallausete osakaal (%)
Substantiivid	
PSV	0%
SS1	12%
etTenTen13	18%
Adjektiivid	
PSV	0%
SS1	14%
etTenTen13	**58%**
Verbid	
PSV	8%
SS1	10%
etTenTen13	**76%**
Adverbid	
PSV	20%
SS1	16%
etTenTen13	76%

Põhjuseks võib olla, et leksikograaf käsitab lauset definitsiooni täiendusena ega lisa infot, mis otseselt ei illustreeri vastava sõna kasutust. Kuid korpuse lausetes peegeldub soov anda rohkem konteksti (näide 5).

(5) Teeme Pere ja Koduga lähemat koostööd ning nii mai, juuni kui juuli numbrist võib lugeda põhjalikke ja asjalikke artikleid lapse seksuaalse arengu teemal ning sellest, kuidas last sel arenguteel toetada. (etTenTen13)

Ilmnes ka, et kõik PSV ja SS1 laused olid süntaktilises mõttes predikatiiviga laused (kas liht- või liitöeldisega). Korpuslausete seas oli palju elliptilisi lauseid (näide 6).

(6) Koht ise päris ilus. (etTenTen13)

Samuti iseloomustab korpuse lauseid suur hulk pärisnimesid ja numbreid (alla joonitud) (näide 7).

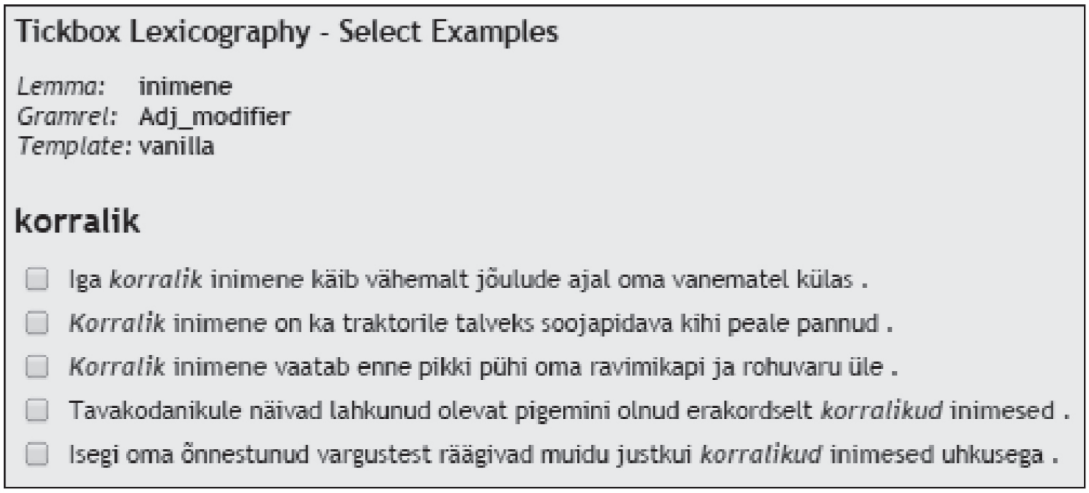
(7) Kosmosesond Voyager 1 on praegu Maast 17,9 miljardi kilomeetri kaugusel – see tähendab, **väga**, **väga** kaugel, kolm kuni neli korda Päikesest kaugemal kui Pluuto. (etTenTen13)

Analüüsi andmetele toetudes töötasime välja GDEX-i eesti mooduli klassifikaatorid, mille põhjal eelistatakse lauseid, mis vastavad järgmistele parameetritele:

- lause algab suure tähega ja lõppeb kirjavahemärgiga;
- lause pikkus on 5–20 sõna;
- lemma ei kordu;
- ei esine sõnu, mis on pikemad kui 20 tähemärki või sisaldavad sümboleid;
- ei esine sõnu, mille sagedus on alla 5;
- lause ei alga sidesõnaga;
- lauses ei esine pärisnimesid, lühendeid, tagasiviiteid *mina*, *sina*, *tema*, *see*, *too* ning adverbe *siin*, *siia*, *siit*, *seal*, *sinna*, *sealt*, *siis*, *seejärel*.

Kuna KOLS on õppeotstarbeline sõnastik, rakendasime lausete genereerimisel musta nimekirja. Nimekirja aluseks oli OÜ Filosofti[15] koostatud loend, milles on sõnad, mida eesti keele speller ei tohi soovitada vigaste ja tundmatute sõnade asendajaks. Nimekirja täiendamiseks uurisime EKSS-i sõnu, mille stiil oli märgitud kui VULG (vulgaarne), HALV (halvustav), KÕNEK (kõnekeelne) või SLÄNG. Musta nimekirja lisasime nt *türa*, *narkots* jne. Peale selle täiendasime nimekirja interneti akronüümide (*omg*, *wtf*, *lol*, *irw*) ning inglis- ja venekeelsete sõimusõnade (*fuck*, *pohui*) ja nende mugandatud variantidega (*fakk*, *pohh*). Lõplik nimekiri koosnes 446 sõnast.

Tulemuseks paranes GDEX-i väljund olulisel määral. Joonisel 7 on esitatud näitelaused kollokatsioonile *korralik inimene*.



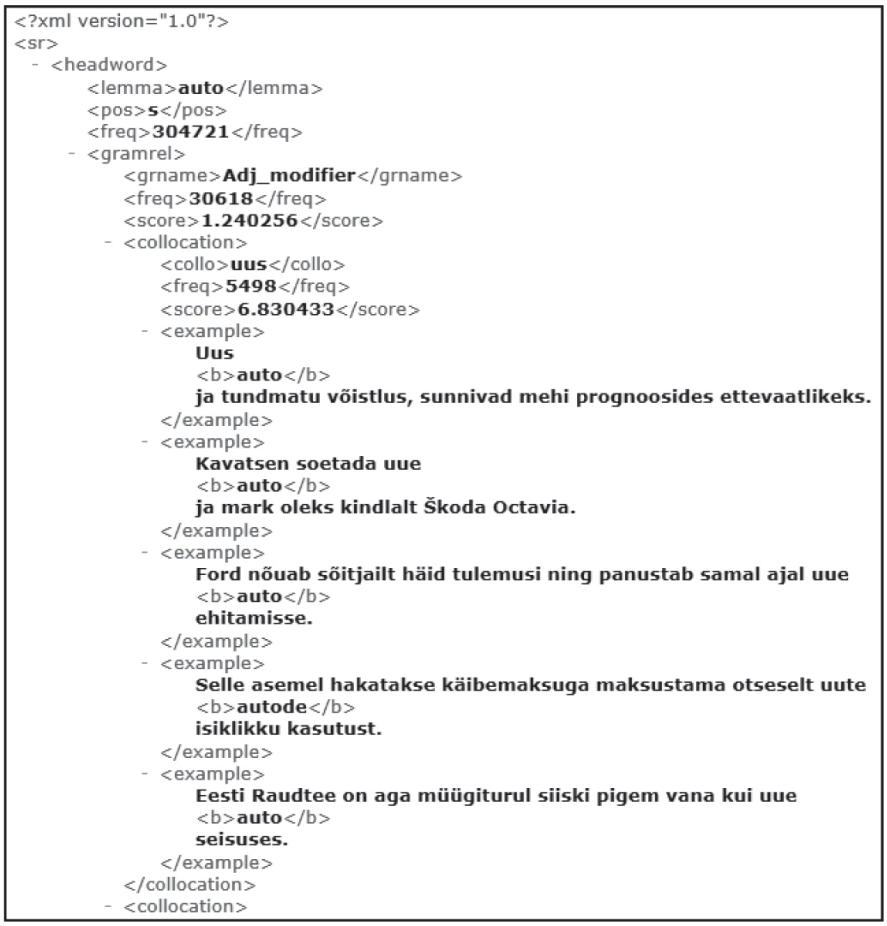
Joonis 7. Kollokatsiooni *korralik inimene* automaatselt valitud näitelaused

4.4. Andmebaasi genereerimine: parameetrid ja tulemused

Andmebaasi genereerimiseks vajalik rakendusprogramm on kirjutatud Pythonis. Andmed saadi korpuspäringusüsteemist XML-faili kujul. Genereerimisel oli kogu märksõnastik jagatud kahte sagedusklassi. Esimese klassi moodustasid märksõnastiku 5000 sagedasemat sõna (minimaalne esinemissagedus korpuses 5057), teise klassi jäid märksõnastiku ülejäänud sõnad. Mõlema sagedusklassi jaoks töötasime välja eraldi parameetrid. Esiteks määrasime kõikide sõnaliikide puhul, missugused kollokatsioonitüübid (grammatilised suhted) automaatselt ekstraheeritakse. Valisime välja 48 grammatilist suhet. Arvestasime vaid neid suhteid, kus kaasmoodustajateks olid substantiivid, adjektiivid, verbid, adverbid ja adpositsioonid. Valikust jäid välja suhted, kus kaasmoodustajaks olid näiteks arvsõnad, asesõnad või pärisnimed. Grammatilise suhte sagedus korpuses pidi olema minimaalselt 10 ja esilduvuse indeks pidi olema positiivne. Sõltuvalt grammatilise suhte tüübist ekstraheerisime iga suhte 5 kuni 20 sagedasemat kollokaati. Lisaks määrasime kollokaatide minimaalse sageduse määra (*minimal frequency*) ja esilduvuse indeksi (*score*). Kollokaadi sagedus korpuses pidi olema vähemalt 10 esimese sagedusklassi sõnade jaoks ja vähemalt 5 teise sagedusklassi jaoks, kollokatsiooni esilduvuse

[15] Autorid tänavad Heiki-Jaan Kaalepit (OÜ Filosoft) loendi eest.

indeks pidi olema positiivne. Kui kollokaat nendele kriteeriumidele ei vastanud, siis seda ei ekstraheeritud. Iga kollokaadi kohta valis programm viis näitelauset. Lausete valikul toetuti GDEX-i väljundile. Joonis 8 illustreerib XML-andmete esitust kollokatsiooni *uus auto* näitel.



Joonis 8. XML-andmete esitus kollokatsiooni *uus auto* näitel

Jooniselt 8 on näha, et automaatselt genereeritud andmebaasi üksused on märksõna (*auto*), märksõna sõnaliik (*s*), märksõna esinemissageduses eesti keele ühendkorpuses (30 472), grammatilise suhte nimetus (*Adj_modifier*), grammatilise suhte sagedus (*freq* 30 618) ja esilduvuse indeks (*score* 1,240256), kollokaat (*uus*), kollokaadi esinemissagedus (*freq* 5498) ja esilduvuse indeks (*score* 6.830433), viis näitelauset. Sel kujul olid andmed imporditud Eesti Keele Instituudi sõnastikusüsteemi EELex. Tulemuseks on andmebaasis 10 939 märksõna, grammatiliste suhete üldarv on 82 678, kollokaate 493 971 ning näitelauseid 2 469 855. Andmebaas sisaldab rohkem üksusi, kui sõnastikus lõpuks esitatakse. See võimaldab toimetamise käigus täpsemat valikut teha.

5. Kokkuvõte

Artiklis analüüsisime tänapäeva korpusleksikograafia üldisi arengutendentse ja uusi suundi Eesti korpusleksikograafias. Automaattuvastamist ja -ekstraheerimist rakendatakse paljude leksikograafiliste infoüksuste puhul. Automaatselt on võimalik luua märksõnastikku, tuvastada kollokatsioone, valida häid näitelauseid, tõlkevasteid jm.

Tutvustasime hiljuti loodud veebikorpust etTenTen13 ning eesti keele ühendkorpuse ülesehitust ja kasutusvõimalusi. Eesti keele ühendkorpus (563 mln sõnet) on hetkel suurim ja žanriliselt mitmekesiseim eesti keele korpus. Ühendkorpuse sisu saab tarkvara Sketch Engine abil analüüsida kogu korpuse ulatuses, allkorpuste (nt ilukirjandustekstid, teadustekstid, ajakirjandustekstid, blogid, foorumid) või domeenide kaupa.

Esitasime valmiva eesti keele kollokatsioonisõnastiku (KOLS) koostamispõhimõtteid ja näitasime, kuidas toimus sõnastiku andmebaasi automaatne genereerimine Sketch Engine'iga. Selleks kasutasime funktsioone: sõnaloend (Word List), sõnavisand (Word Sketch) ja hea näitelause (Good Dictionary Example). Sõnavisandite grammatika viimane versioon 1.6 sisaldab 116 reeglit ja võimaldab tuvastada kõiki KOLS-i jaoks olulisi kollokatsioonirühmi. Lisaks töötasime välja eestikeelsete heade näitelausete automaatseks tuvastamiseks vajalikud parameetrid. Automaatselt genereeritud kollokatsioonisõnastiku andmebaasi maht on 10 939 märksõna, 493 971 kollokaati ja 2 469 855 näitelauset. Järgmine etapp on sõnastiku andmebaasi toimetamine ja täiendamine.

KOLS-i projekti tulemusena valmib mahukas sõnaraamat, mis on suureks abiks eesti keele õppijale. Edaspidi saab sõnastiku andmebaasi kasutada uute leksikograafiliste ressursside loomisel ja eri tüüpi leksikograafiliste veebiportaalide täiendamisel.

Viidatud kirjandus

EKSS = Eesti keele seletav sõnaraamat I–VI. [The Explanatory Dictionary of Estonian.] Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll (Toim.). Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 2009.

Jakubíček, Miloš; Kilgarriff, Adam; Vojtěch, Kovář; Rychlý, Pavel; Suchomel, Vit 2013. The TenTen corpus family. – 7th International Corpus Linguistics Conference CL 2013. Lancaster, 125–127.

Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. [Syntagmatic Relationships of Estonian Content Words in Corpus and Pedagogical Lexicography.] Tallinna Ülikooli humanitaarteaduste dissertatsioonid 32. Tallinn: Tallinna Ülikool. http://e-ait.tlulib.ee/id/eprint/303

Kallas, Jelena; Tuulik, Maria 2011. Eesti keele põhisõnavara sõnastik: ajalooline kontekst ja koostamispõhimõtted. [The basic dictionary of Estonian: The historical context and the principles of compilation.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 59–75. http://dx.doi.org/10.5128/ERYa7.04

Kilgarriff, Adam 2001. Web as corpus. – Proceedings of the Corpus Linguistics Conference (CL 2001), 13 (Special Issue), 342–344.

Kilgarriff, Adam 2013. Using corpora as data source for dictionaries. – Howard Jackson (Ed.). The Bloomsbury Companion to Lexicography. London: Bloomsbury, 77–96.

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

Kilgarriff, Adam; Kovář, Vojtěch; Rychlý, Pavel 2010. Tickbox Lexicography. – S. Granger, M. Paquot (Eds.). eLexicography in the 21st Century: New Challenges, New Applications. Proceedings of eLex 2009, Louvain-la-Neuve, 22–24 October 2009. Louvain-la-Neuve: Presses universitaires de Louvain, 411–418.

Kilgarriff, Adam; Kosem, Iztok 2012. Corpus tools for lexicographers. – S. Granger, M. Paquot (Eds.). Electronic Lexicography. Oxford: Oxford University Press, 31–55.

Kilgarriff, Adam; Rychlý, Pavel; Jakubíček, Milos; Kovář, Vojtěch; Baisa, Vit; Kocincová, Lucia 2014. Extrinsic corpus evaluation with a collocation dictionary task. – Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014).

Kilgarriff, Adam; Rychlý, Pavel; Smrz, 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.

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, Estonia, 17–19.

Kosem, Iztok; Husák, Milos; McCarthy, Diana 2011. GDEX for Slovene. – I. Kozem, 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, 151–159.

Klosa, Annette 2013. The lexicographical process (with special focus on online dictionaries). – H. R. Gouws, U. Heid, W. Schweickard, H. E. Wiegand (Eds.). Dictionaries. An International Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin–Boston: de Gruyter, 517–524.

Langemets, Margit; Mägedi, Marike; Viks, Ülle 2005. Süntaktiline info sõnastikus: probleeme ja väljavaateid. [Syntactic information in dictionaries: Problems and solutions.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 1, 71–98. http://dx.doi.org/10.5128/ERYa1.04

Langemets, Margit; Tiits, Mai; Valdre, Tiia; Voll, Piret 2010. *In spe*: üheköiteline eesti keele sõnaraamat. [A prospective monolingual Estonian dictionary.] – Keel ja Kirjandus, 11, 793–810.

Laufer, Batia 2011. The contribution of dictionary use to the production and retention of collocations in a second language. – International Journal of Lexicography, 24 (1), 29–49. http://dx.doi.org/10.1093/ijl/ecq039

Leech, Geoffrey 2007. New resources, or just better old ones? The Holy Grail of representativeness. Corpus linguistics and the web. – M. Hundt, N. Nesselhauf, C. Biewer (Eds.). Language and Computers, Corpus Linguistics and the Web. Rodopi, 133–149.

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

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

Metslang, Helena; Kibar, Triin 2012. Üldakadeemiline sõnavara. Abivahend eesti keele õppeks kõrgkoolis. [Estonian Academic Vocabulary.] Tallinn: Tallinna Ülikool.

OCDSE 2002 = Oxford Collocations Dictionary for Students of English. Oxford: Oxford University Press, 2002.

Pomikalek, Jan 2011. Removing Boilerplate and Duplicate Content from Web Corpora. PhD thesis. Brno: Masaryk University.

Pomikalek, Jan; Suchomel, Vit 2012. Efficient web crawling for large text corpora. – Proceedings of the 7th Web-as-Corpus workshop, Lyon, France.
PSV = Eesti keele põhisõnavara sõnastik. Jelena Kallas, Mai Tiits, Maria Tuulik (Toim.). Madis Järviste, Kristina Koppel, Maria Tuulik (Koost.). Tallinn: Eesti Keele Sihtasutus, 2014.
Rundell, Michael 2012. How the dictionary was created? http://www.macmillandictionaries.com/features/how-dictionaries-are-written/macmillancollocations-dictionary/ (29.9.2014).
Tiberius, Carole; Schoonheim, Tanneke (ilmumas). The Algemeen Nederlands Woordenboek (ANW) and its lexicographical process. – Vera Hildenbrandt (Ed.). Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks "Internetlexikografie". Mannheim: Institut für Deutsche Sprache.

Jelena Kallas (Eesti Keele Instituut) on eesti keele kollokatsioonisõnastiku töörühma juht. Põhilised uurimisvaldkonnad: õppeleksikograafia, korpusleksikograafia, eesti keele kui teise keele õpetamise metoodika.
Roosikrantsi 6, 10119 Tallinn, Estonia
jelena.kallas@eki.ee

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

Maria Tuulik (Eesti Keele Instituut) on eesti keele kollokatsioonisõnastiku töörühma liige. Põhilised uurimisvaldkonnad: leksikograafia, leksikaalne semantika, tekstilingvistika.
Roosikrantsi 6, 10119 Tallinn, Estonia
maria.tuulik@eki.ee

NEW POSSIBILITIES IN CORPUS LEXICOGRAPHY BASED ON THE EXAMPLE OF THE ESTONIAN COLLOCATIONS DICTIONARY

Jelena Kallas, Kristina Koppel, Maria Tuulik
Institute of the Estonian Language

This article aims to introduce new resources and methods used in Estonian corpus lexicography to create monolingual Estonian dictionaries. Corpora can be used in many ways: headwords list development, grammatical and frequency labels, word sense division, identifying collocations, good dictionary examples, translation equivalents (Kilgarriff 2013). The paper focuses on features offered by Sketch Engine (Kilgarriff et al. 2004), a state-of-the-art lexicographic tool for corpus analysis. For Estonian, Sketch Engine contains different types of corpora, including the recently created 260 million-word web corpus etTenTen13 and the 463 million-word Estonian National Corpus.

Through the example of the Estonian Collocations Dictionary, we analyse how corpus data (headwords, collocations and example sentences) can be automatically extracted from the Estonian National Corpus.

The Estonian Collocations Dictionary contains approx. 10 000 headwords (nouns, adjectives, verbs and adverbs). The various collocates within each headword are grouped according to the lexico-grammatical structure formed by the collocational phrase, and for each collocation one or two example sentences are provided. The main elements needed to develop the algorithm for automatic data extraction are the Sketch Grammar and Good Dictionary Example (Kilgarriff et al. 2008) configurations. The new Sketch Grammar version 1.6 includes all of the lexico-grammatical structures that will be presented in the collocations dictionary. It contains 116 rules in total. For the extraction of dictionary examples, the first version of GDEX for Estonian was developed. Classifiers concerning optimum sentence length, optimum word length, number and type of punctuation marks, word frequency, tokens starting with capital letters, abbreviations etc. were proposed and implemented. The use of classifiers brought significant improvements to the output.

The data was extracted in XML format and imported into the EELex dictionary-writing system, where it will be examined, edited and supplemented by lexicographers. The Estonian Collocations Dictionary will be published in 2018.

Keywords: corpus lexicography, collocations dictionary, corpus query system, dictionary writing system, Estonian