ESUKA – JEFUL 2012, 3 – 2: 57 – 77

LEKSIKOGRAAFILISE TARKVARA SKETCH ENGINE EESTI KEELE MOODUL

Jelena Kallas, Maria Tuulik ja Madis Jürviste
Eesti Keele Instituut

Kokkuvõte. 2010. aasta sügisel alustas Eesti Keele Instituut koos ettevõttega Lexical Computing Ltd. leksikograafilise tarkvara Sketch Engine (Kilgarriff jt 2004) eestikeelse mooduli väljatöötamist. Artiklis kirjeldatakse programmi põhifunktsioone. Põhjalikumalt käsitletakse funktsiooni Word Sketch (ee *sõnavisand*) võimalusi. Tutvustatakse sõnavisandite grammatika koostamise põhimõtteid, vaadeldakse eraldi substantiivide, adjektiivide ja verbide sõnavisandites esitatud süntagmaatilisi seoseid (st grammatilisi ja leksikaalseid kollokatsioone) ning arutletakse mooduli edasiarendusvõimaluste üle. Lisaks analüüsitakse, mil määral saab sõnavisandeid kasutada verbide lausemallide tuvastamisel.

Märksõnad: korpusleksikograafia, konkordantsid, sõnavisand, tesaurus, eesti keel

1. Sketch Engine'i põhifunktsioonid

2002. aastal loodud tarkvara Sketch Engine (SkE)[1] on tekstikorpuste töötlemiseks mõeldud programm, mille põhifunktsioonid on konkordantside koostamine ja nende mitmekülgne töötlemine, korpusest sagedusloendite koostamine, korpusest automaatne sõnavisandite genereerimine ja distributiivne tesaurus (Kilgarriff jt 2004). Kokku on programmis esindatud 42 keele korpused. Tänapäeval kasutavad programmi sõnaraamatute loomisel sellised tuntud kirjastused nagu Oxford University Press, Cambridge University Press, Collins, Le Robert ja Cornelsen Verlag ja seda rakendati ka Dante leksikaalse andmebaasi (Atkins jt 2010) väljatöötamisel. Eestis kasutatakse prog-

1 Sketch Engine vt http://the.sketchengine.co.uk/auth/corpora/. Eesti Keele Instituut sõlmis lepingu 2010. a. sügisel.

rammi praegu üheköitelise seletava sõnaraamatu (Langemets jt 2010) ja eesti keele põhisõnavara sõnastiku (Kallas ja Tuulik 2011) koostamisel.

Programmi eestikeelse mooduli sisendiks on eesti keele koondkorpus[2] (u 250 mln sõnet), mille OÜ Filosoft on morfoloogiliselt märgendanud, osaliselt ühestanud ja osalausestanud. Peale selle on kättesaadav ka eesti lastekeele korpus (u 400 000 sõnet).

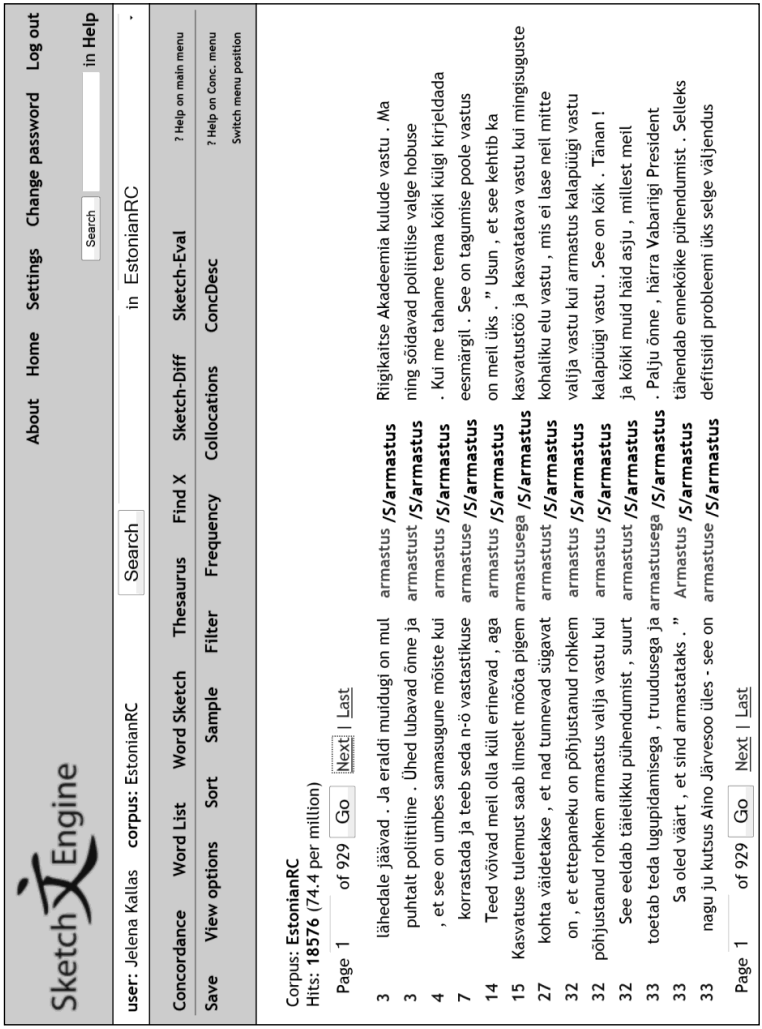
1.1. Konkordantsid ja nende töötlus

Konkordants on sõnavormide loend koos nende tekstilise ümbrusega. Konkordantse kasutatakse mitmes rakenduslingvistika valdkonnas, eelkõige leksikograafias (Atkins jt 2003: 252–253) ja keeleõppes (Kitsnik 2006: 97–98). Adam Kilgarriff (2009) nimetab korpuste tugevuseks just autentset keelt ja ülisuurt näidete arvu. Konkordantsitööriist pakub õppijale suuremat hulka valikuid kui sõnaraamatud, näiteks saab otsida eri parameetrite põhjal ja valida ka viisi, kuidas tulemusi kuvatakse. Sõnaraamatu puhul on õppijal endal valikuvõimalusi paratamatult vähem, kuna leksikograaf on juba ära otsustanud, mida iga sõna puhul teada on vaja.

Sketch Engine on oma olemuselt midagi korpuse ja sõnaraamatu vahepealset. Programmis saab konkordantse koostada sõnavormist, lemmast või fraasist lähtudes või programmi päringukeelt[3] kasutades. Konkordantsi põhisõnale (ingl *keyword*) klõpsates näeb seda sõna lauses koos laiema kontekstiga, mis kuvatakse sõnast vasakul ja paremal. Joonisel 1 on esitatud sõna *armastus* konkordants koos morfoloogiliste märgenditega (sõnaliik, lemma, muutevormi tunnus).

2 Eesti keele koondkorpus vt http://test.cl.ut.ee/korpused/segakorpus/. Korpuse märgendamise kirjeldust vt http://www.filosoft.ee/html_morf_et/morfoutinfo.html

3 vt http://trac.sketchengine.co.uk/wiki/SkE/CorpusQuerying



Joonis 1. Sõna *armastus* konkordants.

Menüüst näeb konkordantside töötlemise võimalusi. Konkordantse saab salvestada (ingl *Save*) eri vormingutes, nt *.txt*- ja *.csv*-vormingusse. Vaate profiili (ingl *View options*) sätete muutmisega saab lisaks põhisõnale vaadata ka kõigi konkordantsides olevate sõnade morfoloogilisi märgendeid, saab määrata kuvatava konteksti pikkust ning konkordantsi ridade arvu. Kuna esialgu oli SkE loodud leksikograafide töövahendiks, siis prooviti parimate näitelausete valimist hõlbustada. Sel eesmärgil lisati Sketch Engine'isse aastal 2008 uus funktsioon nimega GDEX (ingl *Good Dictionary Example*) (Kilgarriff jt 2008). Eestikeelses moodulis on praegu GDEX-i nn *vanilla*-versioon, mis on keelest sõltumatu (olulisteks kriteeriumiteks on lause pikkus ning see, et lause alguses oleks suurtäht ja lõpus punkt, hüüumärk või küsimärk). Mitme keele jaoks on loodud ka keelespetsiifilised GDEX-i rakendused, mis arvestavad ka seda, mil määral sisaldub lauses konkreetse keele põhisõnavara (ingl *core vocabulary*). Vaate profiilis saab määrata, mitut parimat näitelauset soovitakse näha.

Sorteerimise abil (ingl *Sort*) saab konkordantse lahterdada põhisõna või parema ja vasaku naabri järgi. Sagedus (ingl *Frequency*) võimaldab koostada sagedusloendeid lähtuvalt põhisõna lemmast ja sõnavormist, nt saab teada, mis käändes esineb mingi substantiiv kõige sagedamini.

Erinevate statistikute abil (ingl *T-score*, *MI*, *MI3*, *log likelihood*, *min. sensitivity*, *logDice*) saab otsida põhisõna kollokaate (ingl *Collocates*).

1.2. Tesaurus

Distributiivne tesaurus on funktsioon, mille abil saab automaatselt genereerida konkreetse lemma tesauruse. Tegemist ei ole semantilise andmebaasiga, mis keskenduks mõistele ja semantiliste suhete kaudu tema semantilisele väljale (vrd nt Vider 1999), vaid süsteemiga, mis toob statistika põhjal esile sõnu, millel on sarnane grammatiline ja kollokatiivne käitumine. Kuid siiski võib sarnast käitumist põhjustada ka samasse tähendusvälja kuulumine (nt *armastus*, *kirg*, *sõprus* joonisel 2).

armastus

EstonianRC freq = 18576

Lemma	Score	Freq
ilu	0.208	7174
usk	0.201	11245
vabadus	0.196	19110
sõprus	0.194	5372
õnn	0.192	29198
kirg	0.189	5126
rõõm	0.186	17878
tunne	0.184	29587
lootus	0.183	31330

Joonis 2. Lemma *armastus* tesaurus.

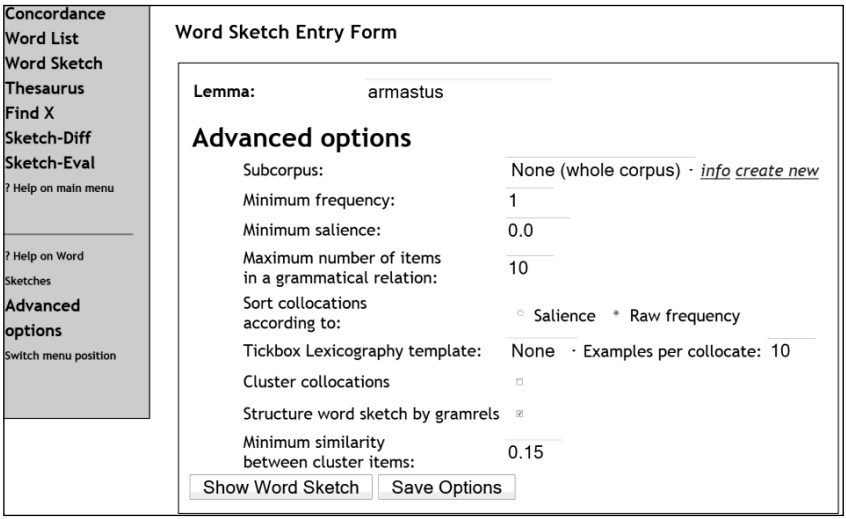
1.3. Sõnavisand

Sõnavisand on automaatne üheleheline kokkuvõte sõna süntaktilisest ja kollokatiivsest käitumisest, mis tugineb korpusele (Kilgarriff jt 2004). SkE-s saab näha ühendit või konstruktsiooni moodustavate sõnade koosesinemiste arvu (ingl *raw frequency*). Kuid sõnadevahelise seose tugevuse mõõtmiseks on rakendatud ka esilduvust (ingl *salience*), mille annab *logDice*-statistik. Kristel Uiboaed (2010: 307) peab statistikute kasutamise eeliseks seda, et peale sõnade koosesinemise võetakse arvesse ka ühendit moodustavate sõnade eraldi esinemise sagedusi.

Sõnavisandi funktsiooni rakendamine eeldab seda, et oleks kirjutatud sõnavisandite grammatika (ingl *Sketch Grammar*), mis määrab, milliseid grammatilisi suhteid (ingl *grammatical relations*) programm otsima hakkab. Süsteem eristab grammatilisi suhteid, mille määrab sõnavisandite grammatika, ja alles seejärel mõõdetakse nende esinemissagedust ja esilduvust.

Sõnavisandite päringu kasutajaliideses toimub otsing lemma kaudu. Laiendatud päringuga (ingl *Advanced options*) on võimalik (vt joonist 3):

- luua isiklik allkorpus (ingl *Subcorpus*), st valida olemasoleva korpuse failidest need, mis hakkavad kuuluma allkorpusesse;
- määrata suhte minimaalne esinemissagedus (ingl *Minimum frequency*);
- määrata grammatilise suhte minimaalne esilduvus (ingl *Minimum salience*);
- määrata ühe kategooria kuvatavate üksuste arv (ingl *Maximum number of items in grammatical relations*);
- sortida kollokaate esilduvuse määra või koosesinemiste arvu järgi (ingl *Sort collocations according to salience/ raw frequency*);
- kasutada funktsiooni "Tickbox Lexicography template", mis pakub eri kollokatsioonide kohta (GDEX-i sätete kohaselt) välja teatud arvu nn parimaid näitelauseid;
- klasterdada kollokatsioone (ingl *Cluster collocations*);
- määrata minimaalne sarnasus klasterdavate üksuste vahel (ingl *Minimal similarity between cluster collocations*).



Joonis 3. Sõnavisandi kasutajaliides.

2. Eesti keele sõnavisandid

Artikli teises osas tutvustame eesti keele sõnavisandite grammatika koostamise põhimõtteid ja näitame sõnaliikide kaupa süntagmaatiliste seoste esitamise printsiipe. Illustreerime ka sõnavisandite võrdlemise funktsiooni kasutust.

2.1. Sõnavisandite grammatika koostamise põhimõtted

Grammatika kirjutamise formalismiks on regulaaravaldised, mis põhinevad sõnaliigil ja muutetunnustel[4]. Lähtuvalt sõnaliigist ja muutetunnustest esitab süsteem sellised grammatilised klassid nagu grammatiline predikaat (öeldis), grammatiline subjekt (alus), grammatiline objekt (sihitis), predikatiiv (öeldistäide), adverbiaal (määrus). Grammatilistest seostest on keskendutud rinnastusseoste ja alistusseoste esiletoomisele. Rinnastusseostest toob süsteem esile konstruktsioonid sidesõnadega *ja/või*, *kui/nagu*. Alistusseostest esitatakse võimalike substantiivi, adjektiivi, verbi ja adverbi rektsioonistruktuuride enamik (vt lähemalt Kallas ja Tuulik 2011: 70).

Grammatilisi suhteid on sõnavisandite grammatikas nelja liiki (vt ka Svensén 2009: 416–418):

a) üheliikmeline suhe (UNARY) on defineeritud ühe konkreetse morfoloogilise kategooria kaudu ning see suhe annab infot ühe kindla grammatilise vormi kohta, nt mis käändes esineb konkreetse noomeni lemma kõige sagedamini;
b) sümmeetriline suhe (SYMMETRIC) puudutab eelkõige rinnastusseoses olevaid ühendeid (eesti moodulis on nendeks *ja/või* ja *kui/nagu* suhted);
c) kaheliikmelised suhted (DUAL) võimaldavad otsida kahe sõna ühendeid ja on määratletud kahe morfoloogilise kategooria kaudu (nt *subjekt, objekt*);

4 St noomeni arvu- ja käändetunnustel, võrdlusastmete tunnustel, verbi arvu-, isiku-, aja-, kõneviisi-, tegumoe- ja kõneliigitunnustel, samuti infiniitsete verbivormide tunnustel.

d) kolmeliikmelised suhted (TRINARY) võimaldavad leida, millised nimisõnad esinevad substantiivi, adjektiivi ja verbi kaassõnafraasides, ja ühendverbide puhul nende laiendeid.

Sõnavisandite grammatika koostamisel on kasutatud ka „Eesti keele formaalses grammatikas" (Roosmaa jt 2001) kirjeldatud kitsendusi. Kokku on programmiga võimalik näidata 38 tüüpi grammatilisi seoseid. Peale selle võimaldab süsteem koostada ühendverbe moodustavate afiksaaladverbide ja verbi laienditega koos esinevate kaassõnade sagedusjärjestusi ning väljendverbide nominaalsete komponentide sagedusjärjestusi tingimusel, et nominaalne komponent on märgendatud kui „X". (X-iga on märgendatud verbi juurde kuuluv sõna, millel eraldi sõnaliigi tähistus puudub, nt *plehku*.)

2.2. Substantiivide sõnavisandid

Substantiivide visandites tulevad esile sellised kategooriad nagu subjekt (mida X teeb?), objekt (mida X-iga tehakse?), adjektiivatribuut (milline X on?) ja eri tüüpi aktantsed atribuudid, mille moodustavad ühelt poolt semantilised käänded ja kaassõnad, teiselt poolt genitiiv. Rektsioonistruktuuridest esinevad visandites käände- (usk *kellesse-millesse)*, kaassõna- (viha *kelle-mille vastu*) ja tegevusnimerektsioon (tahe *mida teha*). SkE abil saab mingil määral vaadelda ka sõna paradigmaatilisi tähendussuhteid, näiteks atributsiooni (*hall, valge hiir*), meronüümiat (*auto rool, mootor, ratas*) ja komplementaarseid vastandusi (*poiss ja tüdruk*), kuna seotud lemmad on tihtipeale kollokatsioonidest väljanopitavad. Allpool on esitatud substantiivide *päike* ja *usk* sõnavisandid.

Päikese sõnavisandis esinevad sellised kategooriad nagu subjekt (*päike paistab, loojub, särab*...), objekt (*päikest nautima, võtma*), adjektiivatribuut (*loojuv, tõusev, lõõskav päike*), genitiivatribuut (*päikese loojumine, kiir, aktiivsus*), aktantne atribuut (*päikese käes*) ja otsitava sõnaga rinnastusseoses olevad substantiivid (*päike ja vihm*, *päike ja kuu*…). Programm

päike () **EstonianRC freq = 16463** (65.9 per million)

object_of	338	3.1
nautima	29	7.19
võtma	76	3.42

subject_of	3106	7.2
paistma	948	10.52
loojuma	147	10.48
särama	116	9.39
kõrvetama	66	9.17
lõõskama	49	8.97

a_modifier	2234	2.3
loojuv	143	10.92
tõusev	271	10.6
lõõskav	110	10.56
ere	164	10.39
kõrvetav	102	10.38

gen_modifies	1615	1.3
loojumine	41	9.64
ultraviolettkiirgus	39	9.27
aktiivsus	91	8.52
kiir	42	8.5
kiirgus	34	8.19

gen_modifier	498	0.4
lõunamaa	25	9.07

ja/või	792	3.2
vihm	52	7.83
tuul	79	7.02
vesi	29	4.02
kuu	40	3.66

noun_seesütlev	230	1.8
käsi	82	4.83

noun_alalütlev	175	1.5
maa	56	3.46

Joonis 4. Substantiivi *päike* sõnavisand.

näitab iga üksuse juures ka koosesinemiste arvu ja esilduvuse määra.

Usu sõnavisandis tulevad esile subjekt (*usk kaob, puudub, aitab*…), objekt (*usku kaotama, sisendama, andma*…), adjektiivatribuut (*hea, kindel, uus usk* …; *katoliku, luteri, vene usk*…), genitiivatribuut (*usu põhimõte, esindaja, puudumine*….), öeldistäide (*usk on oopium*…) ja otsitava sõnaga rinnastusseoses olevad substantiivid (*usk ja lootus*, *usk ja rahvus*…).

Süsteem käsitleb atributiivset suhet kaheliikmelise paarina, seega esitatakse mõned ühendid puudulikult (nt *usu kirik*), kuna neil on veel omakorda laiendid, mis jäävad visandist kõrvale (*luteri usu kirik*). Kollokatsiooni kasutusjuhud koos kontekstiga avanevad koosesinemisarvu alt.

usk *()* **EstonianRC freq = 11245** (45.0 per million)

a_modifier	1623	2.2
hea	319	5.59
kindel	124	6.38
uus	113	3.28
suur	74	2.71
pime	45	7.01
eriline	45	5.6
kristlik	39	7.33

subject_of	582	1.8
puuduma	42	5.57
kaduma	39	5.52
aitama	34	4.87
lubama	23	3.72
tulema	20	1.08
jääma	19	1.62
andma	19	1.45

object_of	473	5.8
andma	47	2.76
sisendama	46	9.69
lisama	33	3.99
kaotama	28	5.04
väljendama	21	6.23
kinnitama	19	2.66
avaldama	16	3.64

a_modifier_comp	49	1.1
suurem	16	2.42

a_modifier_ordinal	72	1.1
teine	71	2.83

adj_modifier_käändumatu	237	8.9
katoliku	81	10.33
luteri	74	11.33
vene	18	3.36
eesti	18	2.93
muhamedi	17	10.62
buda	10	9.71

gen_modifier	1050	1.2
islam	153	10.14
inimene	135	3.55
rahvas	50	4.39
eestlane	24	3.41
juut	19	5.7
moslem	16	7.01
kodanik	15	3.7

gen_modifies	762	0.8
põhimõte	35	4.9
jumal	17	5.05
esindaja	16	2.59
kirik	12	3.2
puudumine	11	3.55
küsimus	11	0.98

olema_noun	80	8.6
uskmatus	11	11.14
oopium	10	9.72

noun_sisseütlev	417	32.6
jumal	66	7.04
õiglus	15	7.14
ime	12	6.38

ja/või	953	5.1
lootus	34	5.82
keel	33	3.71
rahvus	31	6.35
rass	30	8.48
enesekindlus	24	7.79
armastus	23	5.51
teadus	18	5.57

Joonis 5. Substantiivi *usk* sõnavisand.

2.3. Adjektiivide sõnavisandid

Adjektiivifraasi laiendiliikmeks võib olla a) substantiiv(ifraas); b) kaassõnafraas: c) infinitiiv(ifraas); d) adjektiiv(ifraas), e) kvantorifraas; f) adverb(ifraas), g) kõrvallause (EKG 1993: 130). Sõnavisandid näitavad substantiivi, kaassõnafraasi, infinitiivifraasi ja adverbi. Rektsioonistruktuuridest on esitatud käände- (kindel *kelles-milles*), kaassõna- (kade *kelle-mille peale)* ja tegevusnimerektsioon (julge *mida tegema*).

Toome näiteks adjektiivide *hea* ja *rikas* sõnavisandid.

hea *()* **EstonianRC freq = 224446** (898.7 per million)

modifies	146325	7.9
meel	10497	10.85
tulemus	3711	8.87
kolleeg	3244	9.22
võimalus	3094	8.71
sõber	2660	8.88

subj_olema	5466	8.2
vorm	147	6.54
enesetunne	146	8.98
tulemus	116	5.0

da_infinitiiv	1853	0.7
olema	249	3.82
elama	143	6.17
teadma	119	4.53

adverb	62517	5.4
väga	20295	11.36
nii	4993	9.4
päris	2460	9.8
kui	2203	7.74
eriti	1605	8.92

Joonis 6. Adjektiivi *hea* sõnavisand.

Lemma *hea* visandis tulid esile järgmised kategooriad: sagedasemad substantiivid, mille fraasis esineb *hea* adjektiivatribuudina (*hea meel, tulemus, sõber*…). Sellest võib järeldada, et *hea meel* on lemma *hea* sagedasem *substantiiv+adjektiiv*-tüüpi kollokatsioon. Sageduselt teisel kohal on adverbid (*väga, päris, eriti hea*…). Peale selle tulid esile predikatiiv (*enesetunne on hea*, *vorm on hea*, *tulemus on hea*) ja tegevusnimerektsioon (*hea elada*, *hea teada*, *hea olla*).

rikas *()* **EstonianRC freq = 12796** (51.2 per million)

modifies	4546	6.1
riik	665	5.95
inimene	469	5.34
mees	379	5.75
maa	139	4.74
ärimees	72	7.0
pere	67	5.21
naine	65	3.88

subj_olema	242	9.0
riik	13	0.29
ajalugu	7	2.28
keel	6	1.26
linn	6	0.28
maa	6	0.24
vanem	4	1.11
loomaaed	3	4.08

adverb	3872	8.3
nii	684	6.93
väga	426	6.17
kui	262	4.94
ka	189	4.77
ratsa	152	10.27
ainult	121	6.81
vaid	107	6.43

adjective_pp_poolest	57	4778.0
kaalium	3	7.45
antioksüdant	3	7.22
maavara	3	6.93
loodusvara	3	6.76
flavonoidi	2	9.46
B-vitamiin	2	8.87
kiudaine	2	7.51

Joonis 7. Adjektiivi *rikas* sõnavisand.

Lemma *rikas* visandis esinesid järgmised kategooriad: sagedasemad substantiivid, mille fraasis esineb *rikas* adjektiivatribuudina (*rikas riik, inimene, mees...*). Sageduselt teisel kohal on adverbid (*nii, väga, piisavalt rikas*). Lisaks tulid esile predikatiiv (*riik on rikas*, *maa on rikas*) ja kaassõnarektsioon (*rikas mille poolest*).

2.4. Verbide sõnavisandid

Verbide sõnavisandites toob süsteem peale subjekti ja objekti esile ka adverbiaale ehk määrusi. Rektsioonistruktuuridest on esindatud järgmised rektsioonid: objekti- (toetama *keda-mida*), käände- (tutvuma *kellega-millega*), kaassõna- (võitlema

kelle-mille vastu) ja tegevusnimerektsioon (tahtma *mida teha*, jätma *mida tegemata*).

Suur leksikograafiline probleem on aga ühendverbide esiletoomine ning nende süntagmaatiliste seoste selgitamine. Kuna afiksaaladverbid on märgendatud korpuses kui adverbid, siis koostasime EKSS-i ühendverbide põhjal sagedasemate afiksaaladverbide loendi ja lisasime selle sõnavisandite grammatikasse. Tulemuseks on see, et süsteem otsib korpusest ja esitab sagedusloendina ka iga verbi ühendverbe, samuti ühendverbide objekte ja adverbiaale. Teine eesmärk oli tõsta visandites esile, millised kaassõnafraasid laiendavad üht või teist verbi (nt hoolitsema *kelle-mille eest*). Neid suhteid oli võimalik määratleda kolmeliikmeliste kategooriate kaudu.

Toome näiteks verbide *hoolitsema* ja *tooma* sõnavisandid.

hoolitsema *()* **EstonianRC freq = 9376** (37.5 per million)

subject	1133	4.6
riik	78	2.87
firma	31	3.24
mees	27	1.96
valitsus	26	2.12
inimene	24	1.06
ema	22	3.69
naine	22	2.36

adverbial_kaasaütlev	55	4.0
välimus	26	6.45

verb_pp_eest	483	446.9
laps	41	2.81
maja	9	1.3
skoor	8	7.42
loom	8	2.97

adverb	969	2.8
ise	45	6.95
hästi	45	5.56
rohkem	39	5.11
ka	37	2.44
paremini	32	6.16
nüüd	24	3.51
siis	24	2.87

pp_verb	2685	26.1
eest	2574	10.04
kõrval	8	3.69

modifies_past_participle	269	4.1
välimus	43	7.15
aed	18	5.54
käsi	16	2.47
muru	10	5.98

Joonis 8. Verbi *hoolitsema* sõnavisand.

Verbi *hoolitsema* visandis on näha sellised grammatilised klassid nagu subjekt (*kes* hoolitseb), sõltuvusmäärus (hoolitseb *kelle eest*) ja viisimäärus (*kuidas* hoolitseb). SkE tõi välja ka

tud-partitsiibi sagedasemad kollokaadid (*hoolitsetud* *välimus, aed* jne). Nende grammatiliste klasside (uuritava verbi puhul subjekt ja adverbiaalid) alusel võib öelda, et sagedasemad laiendid on: *kes (riik, firma, valitsus)* hoolitseb *kelle-mille eest* (*laste, inimeste eest*) ja *kuidas* (*hästi, korralikult*).

tooma () **EstonianRC freq = 180229** (721.7 per million)

subject	23212	5.5
aasta	243	3.04
tulevik	223	6.42
päev	221	4.81
mees	213	4.8
inimene	198	4.02
naine	117	4.56
elu	114	4.51

object	6608	6.3
näide	840	7.07
selgus	177	8.76
edu	158	7.43
kasu	140	6.42

adv_ühendverb	32416	21.6
kaasa	13557	12.09
välja	6604	9.63
esile	3090	11.03
sisse	2959	10.24
ära	1899	8.4
tagasi	1615	8.62
juurde	659	8.29

adverbial_sisseütlev	2875	13.4
meel	260	7.63
kodu	260	6.43
haigla	102	5.6

verb_pp_ette	536	126.1
avalikkus	149	7.84

adverbial_rajav	413	24.5
vaataja	151	8.32

adverbial_saav	3414	20.8
näide	1571	8.0
põhjus	411	6.91
põhjendus	254	9.32
võrdlus	164	8.88
lisa	141	4.91
ettekääne	126	9.32
eeskuju	125	7.74

adverbial_alaleütlev	6276	15.3
turg	747	8.01
lava	374	8.01
päevavalge	202	9.45
võitja	165	7.11
koha	161	5.08
ilm	160	7.21
ekraan	143	7.65

modifies_past_participle	10399	9.1
näide	334	5.72
tabel	295	8.5
andmed	196	5.25
väide	180	6.49
nõue	115	5.83

Joonis 9. Verbi *tooma* sõnavisand.

Verbi *tooma* visandis on eristatud subjekt (*kes* toob), objekt (*mida* toob), adverbiaalid: 1) latiivne kohamäärus (*tooma kuhu* (*adverbial_alaleütlev*, *adverbial_sisseütlev*)); 2) latiivne valdajamäärus (*tooma kelleni*, *kelle ette* (*adverbial_rajav*; *verb_pp_ette*), sagedasemad on *tooma vaatajani*, *tooma avalikkuse ette*) ja 3) otstarbemäärus (*tooma milleks* (*adverbial_saav*), sagedasem on *tooma näiteks*). Seega verbi *tooma* sagedasemateks laienditeks on subjekt, objekt ja adverbiaalid, mis markeerivad sihtkohta, valdajat ning otstarvet. Peale selle näeb verbi *tooma* visandis ka sagedasemaid ühendverbe: *kaasa tooma, välja tooma, esile tooma* ja *sisse tooma*.

2.4.1. Verbide sõnavisandid kui lausemallide tuvastamise vahend

Kas SkE annab infot otseselt lausemallide kohta või pakub ta vaid sagedasemate laiendite tüüpe? Kas lisaks sõnasemantikale võimaldab visand saada aimu ka konkreetse verbi lausesemantikast? Kas võib oletada, et verbide visandites esile toodud grammatilised klassid (subjekt, objekt, määrus jt) võimaldavad rekonstrueerida verbiga tähistatud tegevussituatsiooni ja selle komponente?

Sõnatähenduse ja tegevussituatsiooni vahekorra üle on arutlenud Huno Rätsep (1978: 237–243), kes on tõdenud, et keeliti võivad tegevussituatsiooni komponendid lauses avalduda situatsiooni kesksest sõnast sõltuvates eraldi sõnades. Sellisel juhul on lause või sõnaühendi keskse sõna tähendus üldisem ja tegevussituatsiooni mõistele kõige lähedasem. Seejuures rõhutab Rätsep (1978: 241), et eesti keeles esitatakse tegevussituatsiooni komponendid pindstruktuuris eraldi sõnadena tegevussituatsiooni mõistet märkiva sõna kõrval.

Teiselt poolt arvatakse, et semantilisi rolle tuleb grammatilistest rollidest ehk lauseliikmetest lahus hoida, sest seost grammatilise käände ja semantilise rolli vahel ei ole ning nimisõna käändevormist ei saa kunagi üheselt järeldada tema semantilist vormi (vt nt Pajusalu 2009: 81–82). Siiski märgib Renate

Pajusalu (*ibid.*), et kuigi semantilistel rollidel ei ole grammatilise vormistusega üksühest seost, pole vorm kunagi tähendusetu.

Allpool proovime lemma *jalutama* näitel rekonstrueerida konkreetse verbi tüüpilise lausemalli, toetudes sõnavisandi statistilistele andmetele. Seejuures eristame lausemalli tuletamisel järgmisi etappe: 1) semantiliste rollide tuvastamine grammatiliste suhete põhjal, mida sõnavisand ette annab; 2) rollide taga olevate entiteetide semantiline analüüs ja kategoriseerimine; 3) lausemalli tuletamine.

jalutama *()* **EstonianRC freq = 9114** (36.5 per million)

subject	612	3.1
mees	50	2.85
inimene	43	1.9
isa	14	3.44
naine	13	1.6
koer	12	4.13
noormees	10	3.89
poiss	7	2.25

adv_ühendverb	1129	16.3
ringi	493	9.8
mööda	135	7.72
läbi	122	5.44
välja	94	3.67
edasi	58	4.88
vastu	51	3.69
üle	32	2.98

pp_verb	933	11.3
mööda	193	8.25
vahel	83	6.19
poole	56	6.7
ääres	53	8.46
peale	49	4.81
läbi	43	3.94
koos	42	4.95

PP_X	183	
verb_pp_mööda	51	513.3
verb_pp_poole	26	75.0
verb_pp_vahel	21	140.2
verb_pp_läbi	20	66.9
verb_pp_juurde	14	35.7
verb_pp_ääres	14	207.4
verb_pp_koos	12	21.6

adverbial_seesütlev	204	11.9
park	40	6.58
mets	21	4.45
linn	19	1.95
vanalinn	11	4.6

adverbial_alalütlev	162	10.3
tänav	36	3.27

adverb	1034	3.7
lihtsalt	54	5.86
kui	51	2.59
väljas	35	8.17
seal	31	4.52
niisama	28	7.49
rahulikult	26	7.71
siis	21	2.68

adverbial_kaasaütlev	104	9.4
koer	20	4.9
laps	10	0.77
poeg	7	3.02

ÜHENDVERB	66	
pp_ringi_ühendverb	45	474.6
pp_välja_ühendverb	21	15.3

Joonis 10. Verbi *jalutama* sõnavisand.

Sõnavisand aitab tuvastada eelkõige verbiga seotud sagedasemad grammatilised klassid, milleks on subjekt (612 esinemisjuhtu) ja eri tüüpi määrused: kaasnemismäärus (104 esinemisjuhtu); lokatiivne kohamäärus (204 esinemisjuhtu); latiivne kohamäärus, kuhu kuuluvad ka *jalutama millegi ääres*-tüüpi määrused; prolatiivne kohamäärus. Seega lähtuvalt esile tulnud grammatilistest klassidest võib selle verbi tegevussituatsioonis eristada järgmisi komponente ehk semantilisi rolle: AGENT, KAASLANE, KOHT, TEE.

Entiteetide analüüs kategooriate järgi näitab seda, et AGENDIKS on *jalutama*-verbi puhul enamasti *inimene*, harvem *loom*; KAASLASEKS on samuti elusolend, kas *loom* või *inimene*; KOHA sagedasemateks täitjateks on substantiivid *park*, *mets, linn ja meri*; TEED märkis sageli substantiiv *tänav.*

Niisiis on statistiliste andmete põhjal *jalutama*-verbi sagedasemad lausemalli komponendid AGENT, KOHT ja KAASLANE, seega võiks sõnastikus välja tuua vastavad verbilaiendid.

2.5. Sõnavisandite võrdlus

SkE lisafunktsioon *Sketch-Diff* võimaldab võrrelda kahe lemma sõnavisandeid. Eelkõige on see funktsioon mõeldud osasünonüümide võrdlemiseks ja eristamiseks. Süsteem toob esile sarnasusi ja erinevusi kahe semantiliselt seotud sõna süntagmaatilises käitumises, st esitab nii sõnade ühised kui ka vaid ühe sõnaga kokkukuuluvad kollokaadid. Näitena esitame katke komplementaarsete vastandite *naine* ja *mees* sõnavisandite võrdlusest (joonis 11).

SkE eristab kollokaate, mis kuuluvad sagedamini kokku lemmaga *naine* (näiteks *rase naine*, *naine sünnitab*, *naist võtma*), ja kollokaate, mis esinevad tihedamini koos lemmaga *mees* (näiteks *tundmatu mees*, *meest süüdistama*, *mees ja naine*). Heledal taustal on atribuudid, mis esinesid enam-vähem võrdselt mõlema lemmaga. Erinevuste esiletoomisel arvestab programm ka sõnade üldist esinemissagedust korpuses.

naine/mees EstonianRC freqs = 179449/307624

naine	6.0	4.0	2.0	0	-2.0	-4.0	-6.0	mees

a_modifier	23848	43752	1.9	2.2
tundmatu	122	905	6.8	9.1
vana	781	1441	8.0	8.6
keskealine	381	708	8.9	9.0
noor	3647	4435	10.0	10.0
rase	348	0	8.8	0.0

subject_of	26160	52914	4.6	5.9
tunnistama	240	619	6.5	7.5
istuma	163	485	6.7	7.7
surema	265	590	7.7	8.1
rääkima	645	929	7.4	7.7
sünnitama	203	0	7.8	0.0

object_of	2034	2789	1.4	1.3
süüdistama	0	116	0.0	8.0
võtma	130	0	4.2	0.0

ja/või	7299	3437	2.3	0.7
naine	0	1753	0.0	8.6
laps	212	0	5.1	0.0
mees	5225	0	9.5	0.0

Joonis 11. Lemmade *naine* ja *mees* sõnavisandite võrdlus.

3. Sketch Engine'i eestikeelse mooduli edasiarendused

Sõnavisandite kvaliteet sõltub vahetult morfoloogilise ühestamise täpsusest. Praegu esineb morfoloogilise märgendamise vigu näiteks sõnaliikide tasandil ning käändsõnade puhul ka käände tasandil. Samuti on üldnimisõnadeks märgendatud isikunimesid (nt Pilv, Jänes, Rebane). Vigu tuleb ka sellest, et sisendkorpuse mitme tõlgendusega sõnade puhul jäi korpuse SkE jaoks ettevalmistamise järgus sisse ainult see tõlgendus, mis oli esikohal. Näiteks esialgu oli verbil *kukutama* kaks vormi: *kukku+tama; kukuta+ma*, kuid SkE-sse on jäänud neist vaid esimene.

Väljundit mõjutab oluliselt ka korpuse tekstide valik (koondkorpuses on praegu u 75% tekstidest ajakirjandustekstid).

Seega tuleb Sketch Engine'it edasi arendades täiendada korpuse sisu ja parandada ühestamise kvaliteeti. Tulevikus võiks korpus olla rohkem tasakaalus ja seeläbi representatiivsem.

Teine arengusuund on sõnavisandite grammatika täiendamine ja selle metakeele (näiteks grammatiliste kategooriate nimetuste) lihtsustamine. Visandite kvaliteeti võiks parandada ka süntaktiliste märgendite kasutamine sõnavisandite gramma-

tika kirjutamisel. See aitaks lahendada nt vaba sõnajärje probleemi. Tuleks püüelda ka semantilise ühestaja kasutamise poole, mida eesti keele jaoks praegu veel ei ole, kuid mida ingliskeelses Sketch Engine'is juba katsetatakse (McCarthy ja Reddy 2010).

Kuna sõnavisandid annavad ülevaate sõna süntagmaatilistest suhetest, saaks Sketch Engine'it edukalt kasutada eesti keele kui teise keele õppimisel ja õpetamisel. Sarnaselt õppesõnastikega toob SkE eksplitsiitselt esile sagedasemad rektsioonistruktuurid ja leksikaalsed kollokatsioonid ning aitab seega õppijat eelkõige just eestikeelse teksti koostamisel. Sõnavisandites näidatud rektsioonistruktuurid peaksid aitama vältida teise keele mõjudest tingitud vigu. Ingliskeelseid sõnavisandeid on juba proovitud automaatselt veebisõnastikuks pöörata (Hvelplund 2011). Tulemuseks on tavakasutajale lihtsasti mõistetav kasutajaliides, mis võimaldab sõnavisandeid vaadata elektroonilise sõnastiku vormingus. Sõnavisandite ja parimate näitelausete valimise funktsiooni kombineerides toimib SkE kollokatsioonide täisautomaatse sõnastikuna. Sellest lähtudes on meie eesmärk ka programmi eestikeelse mooduli veebisõnastikuks pööramine. Sellisel moel saadud n-ö automaatsõnastiku sihtgrupiks oleksid nii eesti keele kui teise ja/või võõrkeele õppijad kui ka tõlkijad, eesti keele uurijad ja kõik teised keelehuvilised.

4. Kokkuvõte

Artiklis tutvustasime leksikograafilise tarkvara Sketch Engine eestikeelse mooduli põhifunktsioone ja rakendamise võimalusi. Konkordantse ja sõnavisandeid saab kasutada eesti keele uurimisel ja kirjeldamisel, (elektrooniliste) sõnastike ja sagedusgrammatikate koostamisel ning eesti keele kui teise ja/või võõrkeele õpetamisel. Sõnavisandites tulevad esile eesti substantiivide, adjektiivide ja verbide sagedasemad süntagmaatilised seosed (rektsioonistruktuurid ja leksikaalsed kollokatsioonid), verbivisandeid saab kasutada ka lausemallide tuletamisel. Sketch Engine on üks esimesi eesti keelele kohandatud korpusleksikograafilisi programme, mille kasutajaskond

ulatub leksikograafide hulgast väljapoole. Kuna korpuspõhisus on tänapäeva keeleteaduse levinumaid arengusuundi, siis on selliste programmide edendamine eesti keele uurimise seisukohast väga oluline.

Aadress:
Jelena Kallas, Maria Tuulik, Madis Jürviste
Eesti Keele Instituut
Roosikrantsi 6
10119 Tallinn, Eesti
E-mail: jelena.kallas@eki.ee, maria.tuulik@eki.ee, madis.jyrviste@eki.ee

Kirjandus

Atkins, Sue, Charles J. Fillmore, and Christopher R. Johnson (2003) "Lexicographing evidence: selecting information from corpus evidence". *International Journal of Lexicography* 15, 251–280.

Atkins, Sue, Adam Kilgarriff, and Michael Rundell (2010) "The DANTE database (Database of analysed texts of English)". *Proceedings of the XII EURALEX International Congress* (Leeuwarden, 6–10 July 2010), 167. Leeuwarden: Fryske Akademy.

EKG 1993 = Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael ja Silvi Vare (1993) *Eesti keele grammatika* II. *Süntaks.* Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Hvelplund, Holger (2011) Using Sketch Engine with IDM's DPS for online dictionaries. Seminari materjalid <https://trac.sketchengine.co.uk/wiki/SKEW-2/Program>. Vaadatud 01.05.2011.

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

Kilgarriff, Adam, Pavel Rychly, Pavel Smrž, and David Tugwell (2004) "The Sketch Engine". *Proceedings of the 11th EURALEX International Congress* (Lorient, 6–10 July 2004), 105–117. Lorient: Université de Bretagne-Sud.

Kilgarriff, Adam, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlý (2008) "GDEX: Automatically finding good dictionary examples in a corpus". *Proceedings of the XIII EURALEX International Congress* (Barcelona, 15–19 July 2008), 105–117 . Barcelona: IULA, Documenta Universitaria.

Kilgarriff, Adam (2009) "Corpora in the classroom without scaring the students". *Proceedings 18th International Symposium on English Teaching.* Taipei.

Kitsnik, Mare (2006) „Keelekorpused ja võõrkeeleõpe". *Eesti Rakenduslingvistika Ühingu aastaraamat* 2, 93–107.

Langemets, Margit, Mai Tiits, Tiia Valdre ja Piret Voll (2010) „In spe: ühekõiteline eesti keele sõnaraamat". *Keel ja Kirjandus* 11, 793–810.

McCarthy, Diana and Siva Reddy (2010) Semantic tagging. Seminari materjalid <https://trac.sketchengine.co.uk/wiki/SKEW-2/Program>. Vaadatud 01.05.2011.

Pajusalu, Renate (2009) *Sõna ja tähendus.* Tallinn: Eesti Keele Sihtasutus

Roosmaa, Tiit, Mare Koit, Kadri Muischnek, Kaili Müürisep, Tiina Puolakainen ja Heli Uibo (2001) *Eesti keele formaalne grammatika.* Tartu: Tartu Ülikooli Kirjastus.

Rätsep, Huno (1978) *Eesti keele lihtlausete tüübid.* Tallinn: Valgus.

Svensén, Bo (2009) *A handbook of lexicography: the theory and practice of dictionary-making.* Cambridge: Cambridge University Press.

Uiboaed, Kristel (2010) „Statistilised meetodid murdekorpuse ühendverbide tuvastamisel". *Eesti Rakenduslingvistika Ühingu aastaraamat* 6, 307–326.

Vider, Kadri (1999) *Sagedamad eesti verbid semantilises andmebaasis.* Käsikirjaline magistritöö. Tartu: Tartu Ülikool. Eesti keele osakond.

Abstract. Jelena Kallas, Maria Tuulik, and Madis Jürviste: Estonian language module of lexicographic software Sketch Engine. The Sketch Engine is a software aiming to analyse corpus data. Its main functions being the compilation and varied analysis of concordances, creating frequency lists, automatic compilation of word sketches and distributive thesauruses.

In the autumn of 2010 the Institute of the Estonian Language started to develop, in collaboration with Lexical Computing Ltd., the Estonian language module for the Sketch Engine software. The Estonian Sketch Engine consists of a 250-million word Estonian Reference Corpus.

The article gives an overview of Estonian Word Sketch grammar compilation principles. The authors demonstrate word sketches for nouns, verbs and adjectives. Also, an insight is given into the possibilities of further development of the Estonian language module.

The corpus-based approach being one of the most widespread tendencies in modern linguistics, it is of utmost importance for Estonian language research to promote the usage of programs such as the Sketch Engine.

Keywords: corpus lexicography, concordance, word sketch, thesaurus, Estonian