TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Monika Muru

# ESINEMISSAGEDUSE MÕJU ÜHENDVERBIDE TÄHENDUSE MOODUSTUMISEL

Bakalaureusetöö

Juhendaja Eleri Aedmaa

TARTU 2018

# Sisukord

# Sissejuhatus

Püsiühendid (ingl k *multiword expression*) on mitmest sõnast koosnevad leksikaalsed üksused (Kühner, Schulte im Walde 2010: 47). Ühese ja laialt aktsepteeritud definitsiooni puudumise tõttu käsitletakse püsiühendeid erinevalt. Semantiliselt peavad mõned uurijad kõiki püsiühendeid mitte-kompositsionaalseteks (ingl k *non-compositional*) ehk ühenditeks, mille tähendus ei ole tema komponentide tähenduse summa. Teistes uurimustes on püsiühendid jaotatud skaalale, mille ühes otsas on mitte-kompositsionaalsed ja teises kompositsionaalsed ühendid ehk ühendid, mille tähendus moodustub tema osiste tähendusest. (Aedmaa 2016: 5)

Püsiühendite uurimine ja tuvastamine on vajalik nii keeletehnoloogidele erinevate keeletehnoloogiliste lahenduste arendamiseks kui ka leksikograafidele sõnastikukirjete loomiseks. (Kühner, Schulte im Walde, 2010: 47) Ka eesti keeles on palju erinevaid püsiühendeid. Süntaktiliselt struktuurilt saab need jagada noomenifraasideks (nt *löök allapoole vööd*), adverbifraasideks (nt *maani täis*), adpositsioonifraasideks (nt *metsa poole*) ning verbi ja tema laiendi moodustatud ühenditeks (nt *läbi saama*). (Kaalep, Muischnek 2009: 159) Selles töös uuritakse ainult eesti keele ühendverbide tähenduse moodustumist.

Mitmete keelte (nt inglise, saksa) püsiühendite tähenduse moodustumisel on ühe faktorina uuritud sageduse mõju nende tähenduse kompositsionaalsusele ja selle määra hindamisele (nt McCarthy jt 2003, Bott, Schulte im Walde 2014). Järeldustena on leitud, et kõige suurema ja väiksema sagedusega ühendite kompositsionaalsust on keerulisem määrata kui keskmise sagedusega ühendite omi (Bott, Schulte im Walde 2014) või on leitud, et sageduse ja kompositsionaalsuse vaheline seos on liiga väike, et selle põhjal kaugeleulatuvaid järeldusi teha (McCarthy jt 2003). Sageduse mõju eesti keele ühendverbide tähenduse moodustumisel ei ole varem põhjalikult vaadeldud.

Töö eesmärk ongi uurida, kas ja mil määral mõjutab eesti keele ühendverbide kompositsionaalsust nende korpuses esinemise sagedus. Varasemate uurimuste põhjal kontrollitakse töös, kas eesti keele ühendverbide puhul kehtib väide, et korpuses sagedamini esinevate ühendverbide kompositsionaalsuse määramine on keerulisem kui harvem esinevate ühendverbide määramine, sest sagedased ühendverbid on tihti mitmetähenduslikud (Bott, Schulte im Walde 2014). Seega on töö uurimisküsimused järgmised.

1. Milline on ühendverbide korpuses esinemise sageduse ja kompositsionaalsuse määra vaheline seos?
2. Kas korpuses sagedalt esinevate ühendverbide kompositsionaalsuse määramine on keerulisem kui harva esinevate ühendverbide kompositsionaalsuse määramine?
3. Kuidas mõjutab ühendverbide polüseemsus nende kompositsionaalsuse määramist?

Töö koosneb kahest suuremast peatükist. Esimeses tutvustatakse ühendverbi, selle tähenduse moodustumist ning tuuakse välja erinevad püsiühendite tuvastamist ja kompositsionaalsuse määramist puudutavad uurimused. Selles peatükis on põhjalik ülevaade Eleri Aedmaa 2017. aastal avaldatud tööst „Exploring Compositionality of Estonian Particle Verbs“ (eesti k „Eesti keele ühendverbide kompositsionaalsuse uurimine“), milles tutvustatud ühendverbide kompositsionaalsuse andmestik on selle bakalaureusetöö aluseks. Töö teine osa annab ülevaate töö tulemustest. Tutvustatud on kasutatud materjali ning välja on toodud ühendverbide kompositsionaalsuse ja sageduse vaheline statistiline analüüs. Lisaks vaadeldakse ühendverbide kompositsionaalsuse ja polüseemsuse vahelist seost.

# 1. Ühendverbi mõiste ja liigitus

Selles peatükis tutvustatakse ühendverbi ja kompositsionaalsuse mõisteid. Tuuakse välja ühendverbide liigitus ning kirjeldatakse nende määratlemist ja varasemat uurimist nii Eestis kui ka mujal. Samuti antakse ülevaade ühendverbide automaatsest tuvastamisest ja sellega seoses tekkivatest probleemidest.

## 1.1. Ühendverbi mõiste

Verbid ehk tegusõnad väljendavad mingit tegevust ning saavad lauses esineda predikaadina (Erelt 2013: 19). Verbid võivad lauses esineda üksikverbina (nt *laulab*), perifrastilise verbina (nt *ette tulema*) või kaksikverbina (nt *mine too*). Kuigi perifrastiliste verbivormide osad kuuluvad eri sõnaliikidesse, moodustavad nad lauses ühtse süntaktilise terviku. Need verbivormid jagunevad omakorda ahel-, väljend- ja ühendverbideks. (EKG II: 18–21, Erelt, 2013) See bakalaureusetöö keskendub ühendverbidele ja nende tähenduse moodustumise uurimisele.

Ühendverbid koosnevad verbist ja afiksaaladverbist ehk abimäärsõnast. Huno Rätsepa järgi on ühendverbi mõiste loonud Elmar Muuk ning selle eristamiseks teistest verbiühenditest on välja toodud neli olulisemat kriteeriumit: ortograafiline, morfoloogiline, süntaktiline ja semantiline. Ortograafilise kriteeriumi järgi on võimalus ühendverbe kirjutada nii kokku kui ka lahku. Morfoloogiline kriteerium ütleb, et ühendverbi moodustavad verb ja abimäärsõna. Süntaktilise kriteeriumi järgi on ühendverb lauses üks tervik. Semantiline kriteerium kirjeldab põhimõtet, et ühendverb kujuneb vaid siis, kui verbile lisatud abimäärsõna tekitab verbiga koos uue mõiste või täiendab verbi tähendust. (Rätsep 1978: 26–27)

Kui verb on ühendverbi sisuline kese, siis afiksaaladverbid lisavad tähendusele nüansi ja väljendavad kas orientatsiooni, perfektiivsust, seisundit või modaalsust.

Orientatsioonilised afiksaaladverbid näitavad suundumis-, paiknemis-, eemaldumis- või kulgemiskohta (vt näiteid 1–4) (EKG II 1993: 20). Erelt (2013: 62) nimetab neid ka kohamäärsõnadeks ehk lokaalseteks afiksaaladverbideks. Selliste abimäärsõnadega moodustatud ühendverbid on näiteks *alla hüppama, kaasa võtma, külge jääma, pärale jõudma, alt minema, läbi lõikama.*

(1) Õun *kukkus* puu otsast *alla*.
(2) Rong *kihutas* õigel ajal *kohale*.
(3) Poiss *tõukas* sõbra *eemale*.
(4) Pruutpaar *sõitis* külast *läbi*.

Perfektiivsust väljendavad afiksaaladverbid märgivad tegevuse piiritletust ehk seda, kas tegevus on lõpetatud või mitte. Neid sisaldavad ühendverbid on näiteks *maha müüma, välja kannatama* ja *läbi lugema*. Levinuimaks perfektiivsusadverbiks on *ära*, mis võib täita nii tegevuse piiritlemise ülesannet (nt *ära tegema*) kui näidata tegevuse orientatsiooni (nt *ära viskama*). (EKG II 1993: 21) Lisaks toob Rätsep (1978: 31) välja abimäärsõnade *tulema* ja *minema* eripära, milles esimene moodustab ühendverbi vaid verbiga *tulema*, kuid teine paljude erinevate liikumisverbidega (vt näiteid 5–6).

(5) Poiss *tuli* sealt *tulema* ja ei vaadanud tagasi.
(6) Koer ehmatas ja *jooksis minema.*

Seisundiadverbid väljendavad mingit seisundit ning jagunevad veel omakorda lokaalseteks ja intralokaalseteks. Esimene rühm sisaldab adverbe, mis näitavad seisundit, milles kõnealune objekt/isik on, ning teine rühm seisundit, millesse siirdutakse. (Rätsep 1978: 47–49) Kõik seisundiadverbid siiski ühendverbe ei moodusta ning need on ühendverbi osad vaid siis, kui moodustub uus tähenduslik tervik või tingitakse lausemall (Erelt 1993: 21–22). Selliselt moodustatud ühendverbid on näiteks *kinni nabima, lahti saama, kokku kukkuma* ja *viltu minema* (Erelt 2013: 64).

Modaalsust väljendavad afiksaaladverbid kujutavad kõneleja suhet väljendatavaga ja vastupidi. Selliste afiksaaladverbidega moodustatud ühendverbid on näiteks *vaja olema/minema, tarvis olema/minema*. (EKG II 1993: 21–22)

## 1.2. Ühendverbi tähenduse moodustumine

Ühendverbid jagunevad „Eesti keele grammatika II“ (edaspidi EKG II) käsitluse järgi tähenduse kujunemise põhjal kahte rühma: ainukordsed (vt näide 7) ja korrapärased (vt näide 8). „Ainukordsed ühendverbid koosnevad piiratud kombinatsioonivõimalustega osistest, mis moodustavad mitte ainult süntaktiliselt, vaid ka semantiliselt liigendamatu terviku, st verbi ja afiksaaladverbi ühend on omandanud uue tähenduse: *peale käima, üle ajama, maha võtma, peale ajama, juurde lõikama, üles ütlema, taga otsima, üle pakkuma, üle pingutama, üles lööma, üles ässitama* jne“ (EKG II 1993: 21). Adverb moodustab verbiga ühe tähendusliku üksuse ning on verbi finiitvormi osa, mitte ei ole verbi seotud laiend, mille kohta saaks eraldi küsimuse esitada (Rätsep 1978: 28). Korrapäraste ühendverbide osised ei ole see-eest piiratud kombinatsioonivõimalustega ning ühe tähendusrühma verbid saavad liituda mingi kindla rühma afiksaaladverbidega. Näiteks liikumisverbid (*minema, jooksma, astuma* jne) liituvad afiksaaladverbidega *alla* ja *eemale*: *alla minema, alla jooksma, alla astuma, alla sõitma, eemale minema, eemale jooksma, eemale astuma, eemale sõitma*. Sellisel juhul jääb mõlemale osisele tähenduslik iseseisvus alles aga tekstis moodustavad nad ühtse süntaktilise terviku. (EKG II 1993: 21)

(7) Ema *otsis* mind *taga*, et koos poodi minna.
(8) Ema *tuli alla* ja läksime poodi.

Arvutilingvistikas kasutatakse püsiühendite (sh ühendverbide) tähenduse moodustamise kirjeldamisel mõistet *kompositsionaalsus* (ingl k *compositionality*) (vt näiteks Kühner ja Schulte im Walde 2010, Bott ja Schulte im Walde 2014, Reddy jt 2011, Aedmaa 2016). Kompositsionaalseteks loetakse püsiühendeid, mille tähendus tuleneb tema komponentide tähenduste summast (näiteks korrapärased ühendverbid, vt näide 9). Mittekompositsionaalsete (ingl k *non-compositional, idiomatic*) püsiühendite tähendus

aga pole tuletatav tema komponentide tähendusest (näiteks ainukordsed ühendverbid, vt näide 10). Kusjuures idiomaatilisuse asemel mitte-kompositsionaalsusest rääkimine on terminoloogiliselt selgem, sest mõnikord loetakse püsiühenditeks vaid idiomaatilisi ühendeid (näiteks Sag jt 2002). Lisaks on välja toodud ka rühm püsiühendeid, mida on nimetatud kompositsionaalseteks idioomideks (Katz, Pitt 2000) ja mis hägustab terminoloogiat veelgi. Siinses töös välditakse seega *idiomaatilisuse* mõistet ning kirjeldatakse ühendverbe kompositsionaalsete või mitte-kompositsionaalsete püsiühenditena lähtudes eespool toodud definitsioonidest.

(9) Auto *sõitis* garaažist *välja*.
(10) Ema *käis* mulle *peale*, et ostaksin endale pesumasina.

Püsiühendite kompositsionaalsuse uurimisel on tavapäraseks saanud, et ühendeid ei jagata kahte gruppi (kompositsionaalseteks ja mitte-kompositsionaalseteks), vaid asetatakse kompositsionaalsuse astme põhjal skaalale (Bannard jt 2003: 65). Seejuures on aga oluline meeles pidada, et mõned sõnaühendid võivad sõltuvalt kontekstist esineda nii kompositsionaalsete kui ka mitte-kompositsionaalsetena, näiteks *välja nägema* (vt lisaks näiteid 11–12) (Aedmaa 2017).

(11) Pille *nägi* enne ballile minemist väga hea *välja*.
(12) Aken oli jääs ning ma *ei näinud välja*.

## 1.3. Ühendverbide ja nende kompositsionaalsuse automaatne tuvastamine

Igapäevases keelekasutuses sellele tõenäoliselt oluliselt ei mõelda, kuid püsiühendite kompositsionaalsuse tuvastamine on oluline näiteks leksikograafidele, et teada, millised ühendid lisada sõnastikesse, ning keeletehnoloogidele, kes peavad erinevate keeletehnoloogiliste programmide loomisel teadma, kas ühendit peaks kohtlema tervikuna või mitte (Kühner, Schulte im Walde 2010: 47). Sellest tulenevalt on

püsiühendite tuvastamiseks ja nende kompositsionaalsuse määramiseks läbi viidud mitmeid uurimusi nii Eestis kui ka mujal maailmas. Järgnevalt tuuakse välja mõned neist.

Erinevate keelte püsiühendeid on palju uuritud. Eriti suurt tähelepanu on saanud inglise keele püsiühendid, kuid vaadeldud on ka näiteks bengali, hindi ja hiina keelte ühendeid (nt Abedin jt 2015; Bhattacharyya jt 2015; Piao jt 2006). Abedin jt (2015) koostasid süsteemi, mis aitab korpustest automaatselt bengali keele püsiühendeid tuvastada ning tulemusena leidsid, et mida suurem on korpus, seda suurem on ka programmi täpsus. Bhattacharyya jt (2015) uurisid hindi keele püsiühendite tuvastamist nende osiste koosesinemise sageduse põhjal WordNeti ja sõnadevahelise koosinuskauguse abil ning leidsid, et hindi keele püsiühendeid aitavad kõige paremini tuvastada WordNeti-põhised lähenemised. Piao jt (2006) kasutasid hiina keele püsiühendite tuvastamiseks inglise keele jaoks loodud statistilist vahendit, mida nad eelnevalt veidi kohandasid. Sarnaselt käesolevale tööle uuriti ka hiina keele püsiühendite tuvastamisel nende sagedust ning leiti, et väga sagedasi püsiühendeid on raskem kindlaks määrata kui harvem esinevaid (Piao jt 2006: 21–22). Adam Goodkind ja Andrew Rosenberg (2015) proovisid leida seoseid püsiühendite tuvastamise ja inimeste trükikiiruse vahel, eeldades, et püsiühendite vahel tehtav paus on väiksem kui teiste sõnade trükkimise vahel tehtav paus. Tulemusena leiti, et paus, mis tehti püsiühendite trükkimise vahel oli tõesti lühem kui ülejäänud sõnade vahel, kuid pauside pikkus olenes suuresti sellest, millisel eesmärgil teksti trükiti. (Goodkind, Rosenberg 2015)

Colin Bannard, Timothy Baldwin ja Alex Lascarides (2003), kes uurisid inglise keele ühendverbe, jõudsid oma uurimuse käigus järeldusele, et püsiühendid tuleb kompositsionaalsuse põhjal klassidesse määramise asemele jagada kompositsionaalsuse skaalale ning Graham Katz ja Eugenie Giesbrecht (2006) kinnitasid hiljem saksa keele püsiühendeid käsitledes sama. Natalie Kühner ja Sabine Schulte im Walde (2010) uurisid saksa keele ühendverbide tuvastamist, rakendades klasterdamismeetodit ja eeldades, et kompositsionaalsemad ühendverbid esinevad rohkem oma põhiverbiga samas klastris. Sel viisil määrasid nad uuritavate ühendverbide kompositsionaalsuse 59% ulatuses.

Eesti keele ühendverb, nagu juba ka öeldud, on samuti püsiühend, seega on ka selle kui ühtse terviku tuvastamine oluline. Eleri Aedmaa (2016) on uurinud, kas ja kuidas on võimalik ühendverbe jagada kaheks – ainukordseteks ja korrapärasteks – ning kas on mõistlikum jagada ühendverbid hoopis kompositsionaalsuse järgi skaalale. Ühendverbide liigitamiseks vaatles Aedmaa ühendisse kuuluvate sõnade koosesinemise sagedust ja kõrvutas tulemusi järgmiste sõnadevahelise seose mõõdikute tööga: t-skoor, vastastikuse informatsiooni väärtus (ingl k *Mutual Information*), hii-ruut-statistik, log-tõepära funktsioon ja minimaalne tundlikkus (*Minimum Sensitivity*). Selgus, et ühendverbide liigitamine kindlatesse klassidesse pole mõõdikuid rakendades võimalik. Ühendverbide kompositsionaalsuse taseme määramiseks kasutas ta koosinuskaugust, sest see meetod on varasemates sõnade tähendust puudutavates uurimustes saavutanud häid tulemusi. *Koosinuskaugus* (ingl k *cosine similarity*) on mõõdik, mida kasutatakse sõnade kontekstivektorivahelise kauguse mõõtmiseks ehk see näitab kahe sõna tähenduse sarnasust nende kontekstide põhjal (Bullinaria, Levy 2007). Ühendverbide kompositsionaalsuse tuvastamisel võib lähtuda hüpoteesist, et „mida lähemal on ühendverbi esitava vektori ja sellesse kuuluva verbi esitava vektori koosinuskauguse väärtus ühele, seda väiksem on vektoritevaheline nurk ja seda sarnasemad on nende ühendverbide ja verbide tähendused“. Tähenduste sarnasus viitab omakorda suuremale kompositsionaalsusele. (Aedmaa 2016: 14)

See uurimus oli eesti keele ühendverbide tähenduse kompositsionaalsuse uurimise kohta esmakordne, kuigi ühendverbide automaatse tuvastamisega on juba ka varem tegeletud. Näiteks tuvastasid Heiki-Jaan Kaalep ja Kadri Muischnek (2002) ühend- ja väljendverbe tekstikorpustest lingvistilisi ja statistilisi meetodeid kasutades ning selgus, et korrektse tulemuse saamiseks tuleb väljundit siiski lisaks ka käsitsi toimetada. Kristel Uiboaed (2010) tuvastas ühendverbe automaatselt murdekorpusest. Selgus, et „statistiku sobivus sõltub kollokatiivse seose tüübist, korpuse suurusest, valdkonnast, keelest jmt“ (Uiboaed 2010: 324). 2013. aastal kasutas Jelena Kallas ühendverbide tuvastamiseks senistele statistilistele meetoditele lisaks ka reeglipõhist lähenemist, mis annab uurijale teatava kontrolli ja mis on arusaadavam kui statistilised meetodid. Sellise lähenemise toimimist kinnitas programmi 70-protsendiline täpsus (Kallas 2013). Kadri Muischnek, Kaili

Müürisep ja Tiina Puolakainen (2014) uurisid ühendverbide tuvastamist automaatse pindsüntaktilise analüüsiga, kasutades leksikoni- ja reeglipõhist strateegiat. Nad leidsid, et leksikonipõhiselt ühendverbe tuvastades saab saagiseks 79,3% ning reeglipõhiselt 97,4%. Eleri Aedmaa (2015) uuris oma magistritöö raames, milliste statistiliste meetodite (t-skoor, vastastikuse informatsiooni väärtus, hii-ruut-statistik, log-tõepära funktsioon, minimaalne tundlikkus, tinglik tõenäosus, ΔP) abil on võimalik eesti keele ühendverbe kõige paremini erinevatest korpustest automaatselt tuvastada. Tulemusena leidis ta, et kuigi t-skoor töötas ajakirjanduskorpusest ühendverbide tuvastamisel hästi, tasub parima tulemuse saamiseks kasutada siiski koos nii sümmeetrilisi kui ka asümmeetrilisi statistikuid, sest tööd mõjutavad nii korpuse suurus kui ka vaadeldavate ühendite arv. (Aedmaa, 2015)

## 1.4. Sageduse ja ühendverbi kompositsionaalsuse seos

Varem on ühendverbide sageduse ja kompositsionaalsuse vahelist seost uurinud näiteks McCarthy, Keller ja Carroll (2003), kes käsitlesid inglise keele ühendtegusõnade kompositsionaalsust. Vaatluse alla võeti Briti Rahvuskorpusest (*British National Corpus*) 4272 ühendtegusõna, millest valiti juhuslikult 100 eri sagedusklassist pärit sõnaühendit ja millele lisati veel 16 valitud sõnaühendit. Töös uuriti ühendverbide kompositsionaalsust erinevate statistikute abil, võrreldes tulemusi inglise keelt emakeelena kõnelevatest inimestest moodustatud testgrupi poolt antud hinnangutega. Lisaks vaadati ka ühendverbi ja selles sisalduva verbi sagedust. Töö käigus leidsid autorid, et verbi ja ühendverbi sagedustel pole mõju ühendverbi kompositsionaalsusele. Verbi sageduse ja kompositsionaalsuse vaheline korrelatsioon oli 0,092 ning ühendverbi sageduse ja kompositsionaalsuse vaheline korrelatsioon -0,096. (McCarthy jt 2003: 76–78)

Saksa keele ühendverbide kompositsionaalsust on uurinud Bott ja Schulte im Walde (2014), kes lisaks kompositsionaalsuse tuvastamisele uurisid ka sageduse mõju selle määramisel. Üheks töö hüpoteesiks oli, et mida sarnasema tähendusega on ühendverb ja sellesse kuuluv verb, seda sarnasemas kontekstis need ka esinevad. Problemaatiliseks

võib aga kujuneda asjaolu, et sagedamini esinevad sõnad võivad olla polüseemsed ehk mitmetähenduslikud. Siiski andmestiku hõredusest (ingl k *sparsity)* lähtudes eeldasid autorid, et sagedamini esinevate ühendverbide kompositsionaalsust on kergem määrata, kuid töö tulemusena leiti, et nii väga vähe kui ka väga palju esinevate ühendverbide kompositsionaalsust määrata on keerulisem kui keskmise sagedusega ühendverbide kompositsionaalsust. Seda ilmestab töös välja toodud tabel (vt tabel 1), mille esimene veerg sisaldab ühendverbide sagedusrühmi ja teine Spearmani korrelatsioonikordajat, mis on leitud vastava sagedusrühma ja kompositsionaalsuse suhtena. Töö autorid peavad sagedamini esinevate ühendverbide keeruka kompositsionaalsuse määramise põhjuseks nende polüseemsust. (Bott, Schulte im Walde 2014) Nende uurimistulemus toetab ka selle bakalaureusetöö eesmärki uurida, kuidas sagedus mõjutab ühendverbide kompositsionaalsuse määra hindamist.

Tabel 1. **Spearmani korrelatsioonikordaja väärtused saksa keele ühendverbi sagedusrühmade ja kompositsionaalsuse vahel** (Bott, Schulte im Walde 2014: 513 järgi)

| **Sagedus** | **Spearmani korrelatsioonikordaja** |
|---|---|
| (2, 5] | 0,16 |
| (5, 10] | 0,27 |
| (10, 18] | 0,26 |
| (18, 55] | 0,59 |
| (55, 110] | 0,25 |
| (110, 300] | 0,06 |
| (300, 6000] | 0,13 |

Tabelist 1 on näha, et madala ja kõrge sagedusega ühendverbid korreleeruvad halvemini ühendverbide kompositsionaalsusega. Kõige madalama sagedusega ühendite rühma sageduse ja kompositsionaalsuse vaheline korrelatsioon on kõigest 0,16 ning kõrgeima sagedusega ühendite korrelatsioon 0,13. Keskmise sagedusega ühendverbide

korrelatsioon on aga 0,59, mida võrreldes teiste tabelis väljatoodud korrelatsioonidega võib pidada väga kõrgeks.

Eleri Aedmaa (2017) on uurinud, kas on võimalik sõnadevahelise seose tugevuse mõõdikuid kasutades tuvastada eesti keele ühendverbide kompositsionaalsust ning asetada need siis vastavalt kompositsionaalsuse skaalale. Selleks kasutas ta eesti keele koondkorpuses eri sagedusega esinevaid ühendverbe. Kompositsionaalsuse määramise alusena kasutas Aedmaa Botti ja Schulte im Walde (2014) hüpoteesi: mida sarnasema kontekstiga on ühendverb ja tema koosseisus olev verb, seda sarnasem on ka nende tähendus, ning väljendas seda koosinuskaugusega (sõnade kontekstivektorite vahelise kauguse mõõdik). Töös toob Aedmaa välja viis kõige kompositsionaalsemat ning viis kõige mitte-kompositsionaalsemat ühendverbi (vt tabel 2) tulenevalt nende koosinuskaugusest. Iga sõnaühendi juures on kirjas ka selle esinemissagedus korpuses ning testgrupi määratud kompositsionaalsuste keskmine (inimestel tuli ühendverbi kompositsionaalsust määrata skaalal ühest viieni, kus üks tähendas madalat kompositsionaalsust ja viis kõrget). Tabelist selgub, et ühendverbi esinemissagedusel võib olla mõju selle koosinuskaugusele – suurema koosinuskaugusega sõnaühendid esinevad korpuses sagedamini, kuid seda pole töös täpsemalt vaadeldud. (Aedmaa 2017: 198–203) Siinne bakalaureusetöö püüabki leida vastust küsimusele, kas ja kuidas mõjutab sagedus ühendverbide kompositsionaalsust.

Tabel 2. **Kõige kompositsionaalsemad ja mitte-kompositsionaalsemad ühendverbid koosinuskauguse põhjal** (Aedmaa 2017: 203 järgi)

| **Koosinuskaugus** | **Ühendverb** | **Sagedus** | **Testgrupi määratud kompositsionaalsuse määr** |
|---|---|---|---|
| 0,44 | *maha müüma* | 8435 | 2,5 |
| 0,43 | *tagasi minema* | 4522 | 4,6 |
| 0,38 | *üle küsima* | 481 | 2,8 |

| | | | |
|---|---|---|---|
| 0,37 | *välja kuulutama* | 13 196 | 3,2 |
| 0,37 | *vastu küsima* | 517 | 3,9 |
| -0,07 | *vahele kukkuma* | 15 | 2,2 |
| -0,09 | *välja saagima* | 59 | 4,4 |
| -0,10 | *välja pilduma* | 19 | 3,6 |
| -0,10 | *peale tungima* | 175 | 2,8 |
| -0,27 | *kokku kiskuma* | 17 | 4,3 |

# 2. Kompositsionaalsuse ja sageduse vaheline seos

See peatükk kirjeldab uurimuse tulemusi. Esmalt antakse ülevaade materjalist – koondkorpusest ning Eleri Aedmaa (2017) loodud ühendverbide kompositsionaalsuse määra andmestikust, mis on osa selle töö põhimaterjaliks. Sellele järgneb analüüs, kus uuritakse kompositsionaalsuse suhet ühendverbide ja selle komponentide sageduse vahel. Eraldi tähelepanu pööratakse ühenditele, mille kompositsionaalsuse hindamine tekitas raskusi. Lisaks analüüsitakse põgusalt ühendverbide kompositsionaalsuse ja polüseemsuse omavahelist seost.

## 2.1. Materjal

### 2.1.1. Eesti keele koondkorpus

Ühendverbide, adverbide ja verbide esinemissagedused on leitud eesti keele koondkorpuses sisalduvatest ajakirjandustekstidest, mida on kokku ca 185 miljonit sõna. Sageduste leidmiseks kasutati korpuse morfoloogiliselt analüüsitud ja ühestatud versiooni, mis saadi Tartu Ülikooli arvutilingvistika uurimisrühma käest.

Koondkorpus on loodud riikliku programmi „Eesti keel ja rahvuslik mälu“ projekti raames ning projekti lõpuks saavutati ka püsitatud eesmärk: korpuses on kokku vähemalt 200 miljonit sõna. (Projekti eesmärgid...) Kõige suurema osa ajakirjanduskorpusest moodustavad Eesti Päevalehe numbrid aastatest 1995–2007 (89,9 miljonit sõna), ajalehe SL Õhtuleht numbrid aastatest 1997–2007 (45,5 miljonit sõna) ja Postimehe numbrid aastatest 1995–2000 (32,9 miljonit sõna). Eesti Ekspressist on 7,2 miljonit sõna ja Maalehest 4,3 miljonit sõna. Kokku sisaldab korpus 13 eri väljaande tekste. (Eesti keele koondkorpus)

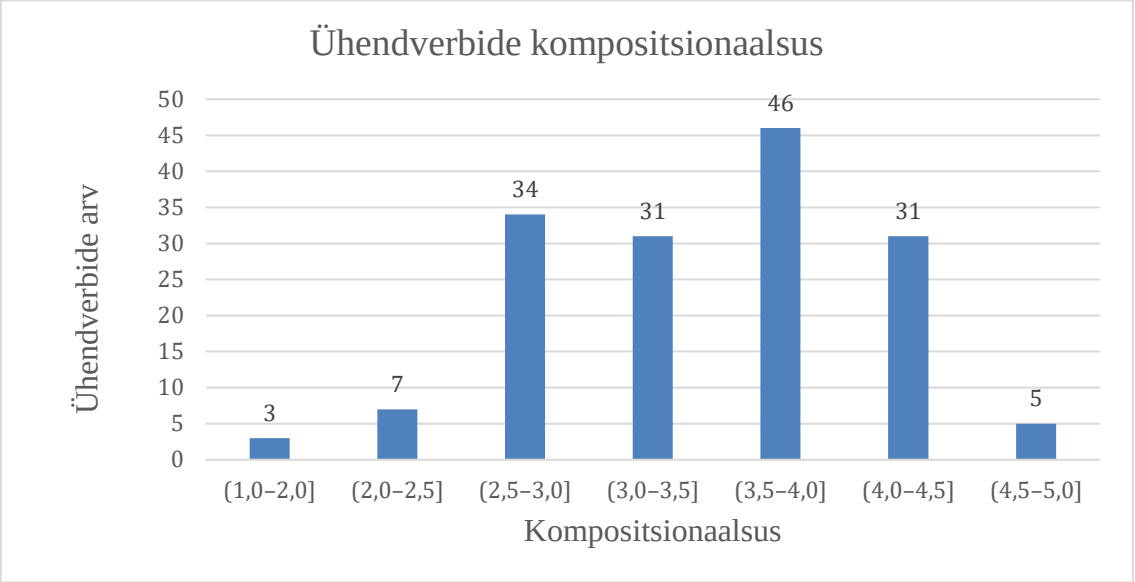
### 2.1.2. Ühendverbide kompositsionaalsuse andmestik

Ühendverbide kompositsionaalsuse määra leidmiseks kasutatakse Aedmaa (2017) uurimuse tarbeks koostatud andmestikku, mis sisaldab infot 211 ühendverbi kompositsionaalsuse kohta. Andmestiku koostamiseks tuvastati koondkorpusest sõnadevahelise seose tugevuse mõõdikuid kasutades automaatselt 1676 ühendverbi (Aedmaa 2016). Seejärel järjestati ühendverbid sageduse järgi kahanevalt, eemaldati väga harva (vähem kui 9 korda) esinevad ühendverbid ja valiti uurimiseks juhuslikult kokku 193 eri sagedusega ühendverbi. Andmestikku lisati ka need ülejäänud 18 ühendverbi, mis juhusliku valikuga andmestikku ei sattunud, kuid mis kuuluvad 20 kõige sagedasema ühendverbi hulka. Seejärel koguti 110 eesti keele kõneleja arvamus nende ühendverbide kompositsionaalsuse määra kohta nii, et vastajal oli võimalus näitelausete põhjal nendes sisalduva ühendverbi kompositsionaalsust hinnata skaalal ühest viieni, kus 1 tähendas madalat kompositsionaalsust (ehk ühendverbi tähendus ei tulene üldse tema komponentide tähendustest) ja 5 kõrget kompositsionaalsust (ehk ühendverbi tähendus tuleneb täielikult tema komponentide tähendustest). Lisaks oli võimalus valida ka kuues variant „ma ei tea“. Skaala oli paarituarvuline, et anda vastajale võimalus valida n-ö kuldne kesktee, mis viitab ühendi mitmetähenduslikkusele. Iga vastaja sai hinnata 21 ühendverbi ning kokku hinnati igat sõnaühendit vähemalt 10 korda. Vastuste kogumine käis veebipõhise küsitluse teel, kus vastajal tuli vastata küsimusele „mil määral moodustub ühendverbi tähendus tema osade tähendusest“.

Andmestik[1] koosnebki valitud ühendverbidest, inimeste hinnangutest ühest viieni, nende põhjal arvutatud standardhälbest, keskmisest kompositsionaalsuse määrast ja vastajate arvust. Kõrge standardhälve tähendab, et hinnangud olid vastajate seas erinevad ning madal standardhälve seda, et hinnangud olid küllaltki sarnased. (Aedmaa 2017) Siinses töös uuritakse eraldi ka neid 54 ühendverbi, mille kompositsionaalsus jäi määramata ehk mille kohta vähemalt üks hindaja märkis vastuseks „ma ei tea“. Skaala, mille põhjal vastuseid anti, oli ühest viieni ning kõigi vastuste põhjal leiti kompositsionaalsuste keskmine ehk kompositsionaalsuse määr (madalaim 1,9 ning kõrgeim 4,67). Joonis 1

[1] Kättesaadav aadressilt https://github.com/eleriaedmaa/compositionality/ (25.05.2018).

illustreerib ühendverbide jaotust nende kompositsionaalsuse määra alusel – ühendverbid jaotati nende kompositsionaalsuse määrade alusel rühmadesse ning joonisel on näidatud ühendverbide arv, mille keskmine kompositsionaalsuse määr vastavasse vahemikku jäi.




Joonis 1. **Ülevaade ühendverbide kompositsionaalsusest**

Jooniselt 1 on näha, et kompositsionaalsuse andmestikus on rohkem kompositsionaalsemaid kui mitte-kompositsionaalsemaid ühendeid. Seega üldistades saab öelda, et eesti keele ühendverbid on pigem kompositsionaalsed ehk sellised, mille tähendus tuleneb tema osiste tähendusest. Samuti on näha ka seda, et väga madala ja vastupidi väga kõrge kompositsionaalsusega ühendverbe esineb andmestikus vähe. Kompositsionaalsusega 1,0–2,0 on ühendverbe kokku kolm (*välja nägema, vastu põrutama, üles kloppima*) ning kompositsionaalsusega 4,5–5,0 on ühendverbe kokku viis (*sisse kutsuma, järele kihutama, tagasi minema, ette jõudma, eemale tõukama*). Kõige rohkem on andmestikus ühendeid, mille kompositsionaalsus jääb vahemikku 3,5–4,0.

Töö eesmärgi saavutamiseks lisati eelnimetatud ühendverbide kompositsionaalsuse andmestikule juurde korpuse põhjal leitud ühendverbide ja nende osiste sagedused. Kuna korpuses ei ole ühendverbid eraldi tähistatud, siis on ühendverbide sagedus umbkaudne. See tähendab et ühendverbiks on loetud iga verbi ja adverbi märgendiga sõnade

koosesinemine ühes osalauses. Sageduse leidmisel on arvesse võetud, et komponentide vahele võivad jääda teised sõnad ning komponentide järjestus pole fikseeritud. Sellel andmestikul põhineb kogu järgnev analüüs.

Ühendverbide sagedused andmestikus varieeruvad väga suurel määral: kõige vähem esineb ühendverbi *ümber riietama,* mida esineb korpuses 9 korda ning kõige sagedasem on ühendverb *vastu võtma,* mida esineb korpuses 35 929 korral. Tabelis 3 on välja toodud andmestiku 10 kõige sagedasemat ühendverbi ning nende kompositsionaalsuse määr (hindajate määratud kompositsionaalsuse määrade keskmine).

Tabel 3. **10 kõige sagedasema ühendverbi kompositsionaalsuse määr**

| **Ühendverb** | **Sagedus** | **Kompositsionaalsus** |
|---|---|---|
| *vastu võtma* | 35 929 | 3,46 |
| *ette nägema* | 28 477 | 3,55 |
| *kinni pidama* | 19 757 | 4,17 |
| *välja andma* | 19 743 | 3,00 |
| *kaasa tooma* | 17 743 | 4,07 |
| *kokku leppima* | 16 009 | 2,80 |
| *ette võtma* | 15 313 | 2,91 |
| *tagasi tulema* | 14 783 | 4,00 |
| *välja kuulutama* | 14 474 | 3,18 |
| *läbi viima* | 14 144 | 2,36 |

Tabelis oleva kümne ühendverbi kompositsionaalsused jäävad vahemikku 2,36–4,17 ning sagedused jäävad vahemikku 14 144 – 35 929. Kõige sagedamini esineva ühendverbi *vastu võtma* kompositsionaalsus on 3,46 ning talle järgneva ühendverbi *ette nägema* kompositsionaalsus 3,55. Selle põhjal võime öelda, et kõige sagedasemad ühendverbid on kompositsionaalsuse skaalal ühest viieni küllaltki keskel, kuid on siiski pigem kompositsionaalsed kui mitte-kompositsionaalsed. Mitte ühegi kõige sagedasema ühendverbi kompositsionaalsuse määr pole alla 1,5 ehk ükski neist ühendverbidest pole

selgelt mitte-kompositsionaalne. Niisamuti pole ühegi ühendverbi kompositsionaalsus üle 4,5 ehk ükski ühendverb pole selgelt kompositsionaalne. Kõige kompositsionaalsem on *kinni pidama* ja kõige mitte-kompositsionaalsem on *läbi viima*. Kokkuvõtlikult saab öelda, et kõige sagedasemate ühendverbide keskmine kompositsionaalsuse määr on veidi üle kolme. Tabelis 4 on välja toodud 10 kõige madalama sagedusega ühendverbi koos nende kompositsionaalsuse määradega.

Tabel 4. **10 kõige madalama sagedusega ühendverbi kompositsionaalsuse määr**

| **Ühendverb** | **Sagedus** | **Kompositsionaalsus** |
|---|---|---|
| *ümber riietama* | 9 | 3,00 |
| *kinni traageldama* | 10 | 4,50 |
| *ümber reastuma* | 12 | 3,64 |
| *läbi kobama* | 13 | 3,82 |
| *üles käänama* | 13 | 3,80 |
| *üles tursuma* | 14 | 4,08 |
| *vahele kukkuma* | 16 | 2,18 |
| *kokku kiskuma* | 18 | 4,30 |
| *kokku valguma* | 18 | 3,90 |
| *välja pilduma* | 20 | 3,64 |

Tabelis 4 on näha andmestiku 10 kõige harvemini esinevat ühendverbi. Nende ühendite sagedus jääb vahemikku 9–20 ja kompositsionaalsuse määr skaalal ühest viieni on vahemikus 2,18–4,5. Kõige vähem esineb ühend *ümber riietama*, mida on korpuses kokku vaid 9 korda. Sellele järgnevad ühendverbid *kinni traageldama* ja *ümber reastuma*, mida on korpuses kokku vastavalt 10 ja 12 korda. Harva esinevate ühendverbide seas on nii kompositsionaalsuse skaala ühele kui ka teisele poolele jäävaid sõnapaare. Veidi enam on siiski suurema kompositsionaalsuse määraga ühendeid, sest alla 3,50 on vaid kaks ühendverbi *vahele kukkuma* ja *ümber riietama*. Kõige kompositsionaalsem ühendverb on *kinni traageldama* ja kõige mitte-kompositsionaalsem *vahele kukkuma*.

Võrreldes kõige sagedasemaid ja kõige harvem esinevaid ühendeid omavahel, tuleb välja, et sagedaste ühendverbide kompositsionaalsuse määrad kipuvad olema mõnevõrra madalamad (ehk mitte-kompositsionaalsemad) kui harvade ühendverbide kompositsionaalsuse määrad. Kümne kõige sagedasema ühendverbi kompositsionaalsuste keskmine on 3,35 ning kõige harvem esinevate sõnapaaride kompositsionaalsuse määrade keskmine on 3,69. Seega, küll väga väikese erinevusega, kuid siiski tuleb kahe tabeli põhjal välja see, et korpuses sagedamini esinevad ühendverbid on mitte-kompositsionaalsemad kui harva esinevad ühendverbid. Siiski ei saa 20 ühendverbi põhjal tulemusi üldistada ning vajalik on täpsem analüüs, mis hõlmab suuremat hulka ühendverbe.

## 2.2. Analüüs

Töös analüüsitakse eraldi kaht gruppi ühendverbe – need, millele kõik hindajad määrasid kompositsionaalsuse taseme skaalal 1–5 ja need, millele vähemalt üks hindaja valis vastuseks „ma ei tea“. Kuigi hindajate ülesanne oli kahtlemata väljakutset pakkuv, nimetatakse selles töös esimesse gruppi kuuluvaid ühendverbe kergesti määratava kompositsionaalsusega ühendverbideks ning teise gruppi kuuluvate ühendverbide kompositsionaalsuse hindamist võib pidada keerukaks. Järgnevalt vaadeldakse, kas ja kuidas mõjutab mõlemasse gruppi kuuluvate ühendverbide ja nende komponentide sagedus ühendi kompositsionaalsust ja kompositsionaalsuse määra hindamist.

### 2.2.1. Kergesti hinnatava kompositsionaalsusega ühendverbid

Selles peatükis analüüsitakse esimest rühma ühendverbe ehk neid, mille kompositsionaalsuse määramisega inimestel probleeme ei tekkinud ehk mis ei saanud kordagi vastust „ma ei tea“. Selliseid ühendverbe oli andmestikus kokku 157. Peatüki lõpus vaadeldakse eraldi neid ühendeid, mille standardhälve viitab võrreldes teiste ühenditega suuremale hindajate vastuste varieerumisele.

Ühendverbi kompositsionaalsuse ja sageduse seose uurimiseks leiti esmalt nendevaheline korrelatsioon. Seejärel arvutati seos kompositsionaalsuse ja adverbide sageduse ning kompositsionaalsuse ja verbide sageduse vahel, mida väljendatakse Pearsoni

korrelatsioonikordajaga. Seose statistilise olulisust vaadeldakse statistilise hüpoteeside kontrollimise teel.

Lineaarne ehk Pearsoni korrelatsioonikordaja mõõdab lineaarset (ehk sirgjoonega kokkuvõetavat) seost kahe arvulise tunnuse vahel. Korrelatsioonikordaja väärtused võivad olla nii positiivsed kui ka negatiivsed ning asuvad vahemikus -1 ja 1. Negatiivne kordaja näitab kahe tunnuse vahelist kahanevat seost ning positiivne vastupidi jällegi kasvavat. Mida lähemale on korrelatsioonikordaja väärtus nullile, seda väiksem on korrelatsioon. Tabelis 5 on välja toodud korrelatsioonimaatriks kompositsionaalsuse ja sageduste vahel. Tärn tähistab seose statistilist olulisust.

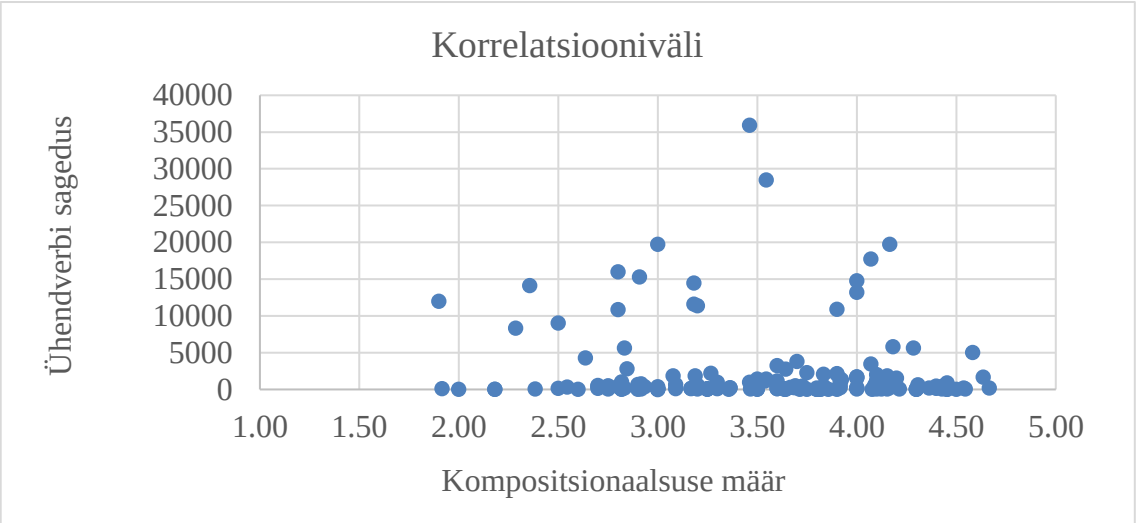
Tabel 5. **Korrelatsioonimaatriks 157 ühendverbi põhjal**

| | **Kompositsionaalsuse määr** | **Adverbi sagedus** | **Verbi sagedus** | **Ühendverbi sagedus** |
|---|---|---|---|---|
| **Kompositsionaalsuse määr** | 1 | | | |
| **Adverbi sagedus** | 0,085 | 1 | | |
| **Verbi sagedus** | 0,046 | -0,103 | 1 | |
| **Ühendverbi sagedus** | -0,085 | 0,059 | 0,548* | 1 |

Tabelis 5 esitatud korrelatsioonikordajad kompositsionaalsuse määra ja sageduste vahel viitavad kõik väga nõrgale seosele. Seos kompositsionaalsuse määra ja ühendverbi esinemissageduse vahel on -0,085. Adverbi sageduse ja kompositsionaalsuse määra vaheline korrelatsioonikordaja 0,085 ning verbi sageduse ja kompositsionaalsuse määra seos 0,046. Lisaks on tabelis ka seosed ühendverbi sageduse ja tema osiste sageduste vahel, kus ilmneb tugev seos ühendverbi ja selles sisalduva verbi sageduste vahel (0,548). See tuleneb töö esimeses peatükis välja toodud väitest, et verb on ühendverbi sisuline kese (vt ptk 1.1) ning sagedased verbid moodustavad ka sagedamini esinevaid ühendverbe.

Tulemuste statistiliste hüpoteeside kontrollimisel ilmnes, et ainult ühendverbi ja verbi sageduste vahelise seose kohta saame väita, et tulemus on statistiliselt oluline nivool 0,05 (olulisustõenäosus $p<0,05$). Teiste seoste puhul on p-väärtus suurem kui 0,05 ning seos ei ole statistiliselt oluline. Võttes arvesse esitatud korrelatsioonikordajaid ja hüpoteeside kontrolli, saab öelda, et vastu tuleb võtta nullhüpotees ehk seost sageduse ja kompositsionaalsuse vahel võib mitte olla. Seose puudumist ega olemasolu aga ei saa praeguse andmestiku peal tõestada ning vajalik on suurema andmestiku uurimine.

Lisaks korrelatsioonikordajatele illustreerib tulemusi ka joonis 2, kus on ühendverbide kompositsionaalsuse ja sageduse põhjal esitatud hajuvusdiagramm. Jooniselt on näha, et kõrge sagedusega ühendverbid on keskmise kompositsionaalsuse määraga, mis võib näidata, et neil ühenditel ei ole ühte selget põhitähendust, mille kõik hindajad aluseks võtsid. Hinne *3* võib tähendada ka seda, et hindajad püüdsid viidata sagedaste ühendverbide kalduvusele olla mitmetähenduslikud. Selle, et sagedased ühendverbid on polüseemsemad kui need, mis korpuses harva esinevad on näiteks välja toonud Stefan Bott ja Sabine Schulte im Walde (2014: 514).




Joonis 2. **Korrelatsiooniväli ühendverbide sageduse ja kompositsionaalsuse määra vahel**

Edasiseks analüüsiks jagati kõik 157 vaatluse all olevat ühendverbi nende sageduse põhjal kolme rühma: väikese sagedusega (esineb korpuses 9–100 korda), keskmise sagedusega

(esineb korpuses 101–1000 korda) ja kõrge sagedusega (esineb korpuses 1001–36 000 korda). Rühmades on ühendverbe kokku vastavalt 46, 60 ja 51. Tulemusi illustreerib tabel 6, kus on näha kolme ühendverbide sagedusrühma seos kompositsionaalsuse määraga. Lisaks on tabelis ka iga sagedusgrupi keskmine kompositsionaalsus.

Tabel 6**. Korrelatsioonid ühendverbide sagedusrühmade ja kompositsionaalsuse määra vahel**

| | **Sagedusgrupi keskmine kompositsionaalsus** | **Korrelatsioon sageduse ja kompositsionaalsuse vahel** |
|---|---|---|
| **Väikese sagedusega ühendverbid** | 3,54 | -0,048 |
| **Keskmise sagedusega ühendverbid** | 3,45 | 0,023 |
| **Kõrge sagedusega ühendverbid** | 3,53 | -0,233 |

Keskmised kompositsionaalsuse määrad on sagedusgrupipide võrdluses sarnased – väikese sagedusega ühendverbidel 3,54, keskmise sagedusega ühenditel 3,45 ning kõrge sagedusega ühenditel 3,53. Väikese ja keskmise sagedusega ühendverbide korrelatsioon sageduse ja kompositsionaalsuse vahel on peaaegu olematu, kuid kõrge sagedusega ehk korpuses 1001 ja enam korda esinevate ühendverbide korrelatsioon kompositsionaalsusega on veidi tugevam, täpsemalt -0,233. Negatiivne kordaja väärtus viitab negatiivsele seosele ehk sageduse suurenemine toob kaasa kompositsionaalsuse vähenemise. Samas on tegemist üsna madala korrelatsiooniga ning nagu jooniselt 2 näha on, siis sagedasete ühendverbide seas on nii kõrge kui ka madala kompositsionaalsusega ühendeid. Siiski vaatlus, et sagedased ühendverbid on keskmiselt mõnevõrra mitte-kompositsionaalsemad kui harvad ühendverbid, selgus ka eelnevalt kõige sagedasemaid ühendverbe analüüsides (vt ptk 2.1.2).

Tulemuste statistiliste hüpoteeside kontroll selgitas, et seoste puhul on p-väärtused suuremad kui 0,05, seega seos ei ole statistiliselt oluline ja tulemus kehtib vaid siin töös kasutatud andmestiku kohta. Kuna selle andmestiku peal ei saa seose olemasolu ega ka puudumist tõestada ning tabelis 6 esitatud korrelatsioonikordajad viitavad võrreldes teiste sagedusrühmadega kõrge sagedusega ühendverbide tugevamale seosele kompositsionaalsuse määraga, siis tasub kindlasti jätkata ühendverbide kompositsionaalsuse määra ja sageduse vaheliste seoste uurimisega (suurema andmestiku peal).

Lisaks kompositsionaalsusele ja sagedusele analüüsitakse töös lähemalt ka andmestikus olevate ühendverbide kompositsionaalsuse hinnangute põhjal arvutatud standardhälbeid. 157 ühendverbi seas on ühendverbe, mille standardhälve oli võrreldes teistega kõrgem. Mida kõrgem on standardhälve, seda suurem on erinevus hinnangute seas, mis inimesed ühendverbide kompositsionaalsuse määramisel andsid. Teisisõnu inimeste vastused nende ühendverbide hindamisel varieerusid rohkem ning üks ja sama ühendverb võis ühe hindaja poolt kompositsionaalsuse määraks saada ühe, kuid teise hindaja poolt jällegi viie. Täpsem info hindamisskaala kohta on leitav peatükis 2.1.2. See, kuhu tõmmata piir kõrge ja madala standardhälbe vahele, pole selge, kuid siinses töös vaadeldakse neid ühendverbe, mille standardhälve on kõrgem kui 1,5. Selliseid ühendverbe oli kokku 15.

Sarnaselt on inglise keele nimisõnaühendeid uurinud Reddy jt (2011) kasutanud testgrupi poolt saadud hinnangute standardhälvet, et analüüsida saadud tulemuste sobivust edaspidiseks uurimiseks. Nende uuritud 90 püsiühendist oli standardhälve suurem kui 1,5 15 ühendil ning uurijad eeldasid, et selle põhjuseks oli kas ühendi polüseemsus või hindajate erinev subjektiivne arvamus. (Reddy jt 2011)

Tabelis 7 on välja toodud kõik 15 andmestikus olnud ühendverbi, mille standardhälve oli kõrgem kui 1,5. Iga ühendi juures on kirjas tema standardhälve, ühendverbis sisalduva adverbi ja verbi sagedused ning ühendverbi enda sagedus. Lisaks on iga ühendi juurde lisatud tema erinevate tähenduste arv „Eesti keele seletavas sõnaraamatus“ (edaspidi

EKSS), mis aitab hiljem määratleda seda, kas ühendverb on pigem ühe- või mitmetähenduslik (EKSS).

Tabel 7. **Kõrge standardhälbega ühendverbid**

| **Ühendverb** | **Standard-hälve** | **Adverbi sagedus** | **Verbi sagedus** | **Ühendverbi sagedus** | **Tähendused** |
|---|---|---|---|---|---|
| *ette heitma* | 1,85 | 124 146 | 18 862 | 5638 | 1 |
| *ümber riietama* | 1,83 | 28 792 | 1574 | 9 | 1 |
| *välja loosima* | 1,73 | 322 547 | 2615 | 1414 | 1 |
| *läbi kaaluma* | 1,64 | 99 135 | 21 242 | 412 | 1 |
| *ümber reastuma* | 1,63 | 28 792 | 159 | 12 | 1 |
| *ümber mõtlema* | 1,60 | 28 792 | 122 229 | 693 | 1 |
| *üles peksma* | 1,60 | 70 433 | 9117 | 34 | 3 |
| *välja pakkima* | 1,60 | 322 547 | 4021 | 48 | 1 |
| *lahti pääsema* | 1,58 | 37 269 | 65 253 | 362 | 3 |
| *ette võtma* | 1,58 | 124 146 | 350 220 | 15 313 | 4 |
| *alla laadima* | 1,56 | 63 131 | 3985 | 362 | 1 |
| *alla käima* | 1,56 | 63 131 | 190 363 | 791 | 1 |
| *lahti siduma* | 1,51 | 37 269 | 65 931 | 152 | 1 |
| *ümber ristima* | 1,51 | 28 792 | 2565 | 65 | 1 |
| *vastu raiuma* | 1,50 | 82 284 | 4618 | 55 | 1 |

Tabelist 7 on näha, et kõrge standardhälbega ühendverbid on korpuses küllaltki harva esinevad. Kõige sagedasem ühend nende seas on *ette võtma*, mida esineb korpuses 15 313 korda, kuid millel on samas ka EKSSi põhjal neli erinevat tähendust. Seega on sage korpuses esinemine ka loogiline, sest eri tähendusi kasutatakse erinevates kontekstides. Sageduselt teine ühendverb on *ette heitma*, mida esineb korpuses 5638 korda ning sageduselt kolmas on ühendverb *välja loosima*, mida on korpuses 1414 korda. Ülejäänud ühendverbid esinevad korpuses vähem kui 1000 korda. Väga selgelt on tabelist 7 näha ka see, et kõige kõrgema standardhälbega ühendverbid on ühetähenduslikud, mis on huvitav

tulemus. Nimelt võib eeldada, et just ühe tähendusega ühendverbid on need, mis võiks saada hindajatelt sarnaseid hinnanguid.

Ühendverb *ette heitma* on tabelis kõige kõrgema standardhälbega ühend, mis tähendab, et selle määramisel olid inimesed kõige rohkem eri arvamusel. EKSSi põhjal on sellel aga vaid üks tähendus, milleks on *etteheiteid tegema*. Ühetähendusliku ühendverbi keerukas kompositsionaalsuse määramine on üllatav ning ei ühti töö alguses püstitatud väitega, mis ütles, et kompositsionaalsuse määramine on keeruline mitmetähenduslikel ühendverbidel. Üheks põhjuseks, miks ühendverbide standardhälve on neil ühetähenduslikel ühenditel kõrge, peab töö autor ühendi osiste sagedust. Bott ja Schulte im Walde (2014) uskusid, et sagedaste ühendverbide kompositsionaalsuse määratlemine on keeruline, sest need on polüseemsed, kuid sagedased adverbid ja verbid võivad samuti olla polüseemsed ning see muudab ühendverbi kui terviku kompositsionaalsuse määramise keeruliseks.

Näiteks näeme tabelist 7, et ühendverbi *ette heitma*, mis oli kõrgeima standardhälbega, adverb on küllaltki sage ning esineb korpuses kokku 124 146 korda (see on andmestikus sageduselt neljas adverb). EKSSi põhjal selgub ka, et nii adverb *ette* (tähendusi 12) kui ka verb *heitma* (tähendusi 5) on ise ka küllaltki polüseemsed sõnad.

Lisaks uuriti lähemalt ka ühendeid *ümber riietama* ja *välja loosima*, mis on saanud samuti kõrge standardhälbe. Adverbid *ümber* (tähenduste arv 10) ja *välja* (tähenduste arv 7) on ootuspäraselt ka väga polüseemsed, kuid verbid *riietama* ja *loosima* pigem mitte. Seejuures on adverb *välja* ka oma sagedusega (322 547) kogu andmestiku kõige sagedasem adverb. Adverb *ümber* esineb korpuses 28 792 korda, mis on võrreldes eelnevate sagedustega tegelikult madal. Samuti on madal verbi *riietama* sagedus – 1574. Sellest järeldab töö autor, et üheks kõrge standardhälbe ehk inimeste erimeelsuse põhjuseks võib olla sage ja polüseemne ühendverbi osa. Polüseemne sõna võib erinevatel inimestel ühendverbi analüüsides tekitada erinevaid tähendusseoseid ning seetõttu hinnatakse ühe sõna erinevaid tähendusi, mis saavad omakorda erineva hinnangu.

Ühendverbide kompositsionaalsuse, standardhälbe ja polüseemsuse vahelise seose analüüs on leitav peatükist 2.2.3.

Kokkuvõtlikult ei saa kasutatud andmestiku põhjal öelda, et ühendverbide ja selle komponentide sagedus mõjutab ühendverbide tähenduse kompositsionaalsust. Niisamuti pole võimalik kinnitada, et seost sageduse ja kompositsionaalsuse vahel pole. Kõrge standardhälbega ühendverbide tähenduste arv aga lubab oletada, et ühendverbide kompositsionaalsuse hindamise teeb keerukaks komponentide polüseemsus.

### 2.2.2. Raskesti hinnatava kompositsionaalsusega ühendverbid

Siin peatükis võetakse vaatluse alla ühendverbid, mille kompositsionaalsuse määramisel inimesed kindlat otsust teha ei suutnud ehk mis said vähemalt üks kord väärtuseks „ma ei tea“. Neid ühendverbe oli kokku 54, mis tähendab, et raskesti hinnatava kompositsionaalsusega ühendverbe oli andmestikus vähem kui neid, mida oli kerge hinnata (157). Järgnevalt analüüsitakse, millistel põhjustel võisid just need ühendverbid hindajatele raskusi valmistada keskendudes eelkõige sageduse mõjule.

Nende ühendverbide seas oli nii sagedasi kui ka harva esinevaid ühendverbe. Näiteks sai üks kord määratluseks “ma ei tea” ühendverb *välja tulema,* mida esines korpuses 32 454 korda ning kaks korda ühendverb *maha tantsima*, mida esines korpuses kõigest 14 korda. Mõlema ühendverbide rühma (kergesti ja raskesti määratavate kompositsionaalsustega ühendverbid) vahel on harvu ja sagedaid ühendverbe küllaltki võrdselt. Esimeses rühmas on sagedusega 1000 või rohkem ühendverbe kokku 33% ning teises rühmas 31%.

Iga ühendverbi hindas minimaalselt kümme inimest ning ükski vaadeldav ühendverb ei saanud määratluseks “ma ei tea” kõigilt kümnelt hindajalt. Maksimaalne “ma ei tea” hulk ühe ühendverbi määratlemisel on kolm. Tabelis 8 on välja toodud 14 ühendverbi, mis said kompositsionaalsuse määramisel kõige rohkem (ehk kaks või kolm korda) vastuseks „ma ei tea“. Ühendid on järjestatud sageduse põhjal. Lisaks on tabelis ühendverbi osade sagedused, vastuse „ma ei tea“ hulk ja tähenduste arv EKSSi põhjal.

Tabel 8. **14 kompositsionaalsuse määratlemisel kõige keerulisemaks osutunud ühendverbi**

| Ühendverb | Sagedus | Adverbi sagedus | Verbi sagedus | „Ma ei tea“ | Tähendused |
|---|---|---|---|---|---|
| *ära tegema* | 13 828 | 191 640 | 650 696 | 2 | 3 |
| *välja võtma* | 7520 | 322 547 | 350 220 | 2 | 3 |
| *üle käima* | 3380 | 145 907 | 190 363 | 2 | 3 |
| *üles võtma* | 3332 | 70 433 | 350 220 | 2 | 3 |
| *ette tegema* | 1594 | 124 146 | 650 696 | 2 | 3 |
| *maha ajama* | 686 | 80 310 | 50 395 | 3 | 4 |
| *üle trumpama* | 476 | 145 907 | 591 | 3 | 1 |
| *maha saagima* | 316 | 80 310 | 1403 | 2 | 1 |
| *sisse õnnistama* | 252 | 57 712 | 1621 | 2 | 2 |
| *järele laskma* | 108 | 19 230 | 82 027 | 2 | 1 |
| *üles vuntsima* | 67 | 70 433 | 155 | 2 | 1 |
| *välja nutma* | 37 | 322 547 | 6655 | 2 | 1 |
| *kokku trehvama* | 33 | 164 868 | 515 | 3 | 1 |
| *maha tantsima* | 14 | 80 310 | 8621 | 2 | 1 |

Need ühendverbid, mille kompositsionaalsuse määramisel hindajatel kõige rohkem raskusi tundus olevat ehk mis said hinnangu “ma ei tea” kolm korda, olid *kokku trehvama, üle trumpama* ja *maha ajama*. Tabelis 8 olevate ühendverbide sagedused jäävad vahemikku 14–13 828 ning rohkem oli just harva esinevaid ühendeid. Arvestades, et keskmine sagedus korpuses selle rühma (raskesti hinnatava kompositsionaalsusega) ühendverbide juures oli 2527, võib öelda, et kõige rohkem segadust tekitanud ühendverbid on korpuses küllaltki harva esinevad.

Kuus kõige sagedasemat ühendverbi, mis said määratluse „ma ei tea“ kaks või enam korda olid *ära tegema, välja võtma, üle käima, üles võtma, ette tegema* ja *maha ajama.* Tabelist 8 on näha, et nendel ühenditel on EKSSi põhjal rohkem tähendusi kui teistel.

Samas on kolm korda väärtuse „ma ei tea“ saanud ka ühendverbid *üle trumpama* ja *kokku trehvama*, millel on üks tähendus.

Ühe põhjusena, miks ühetähenduslikud ühendverbid on raskesti määratletavad, tõi töö autor eelnevalt (vt ptk 2.2.1) välja selle, et ühend sisaldab polüseemset adverbi või verbi. Ühendverbide, mis said määratluse „ma ei tea“ kolm korda, osad *kokku* (sagedus 164 868), *üle* ( sagedus 145 907) ja *ajama* (sagedus 50 395) on tabeli 8 põhjal väga sagedased ja samuti ka väga polüseemsed. EKSSi põhjal on adverbidel *kokku* 11 ja *üle* 16 tähendust ning verbil *ajama* ka 16 tähendust. Andmestikus on ka adverbe ja verbe, mis on määratluse „ma ei tea“ saanud mitme ühendverbi koosseisus ning nendeks on *välja, üle, üles, maha, tegema* ja *võtma.* Need sõnad on andmestikus ka ühed sagedasemad ning EKSSi põhjal ka ühed polüseemsemad. Näiteks on adverbidel *välja* 7, *üles* 21 ja *maha* 16 tähendust ning verbidel *tegema* ja *võtma* vastavalt 18 ja 17 erinevat tähendust. Lisaks on adverb *välja* kogu andmestiku kõige sagedasem ning verb *tegema* andmestikus sageduselt teine verb. Analüüsi tulemusena võib järeldada, et ühendverbide kompositsionaalsuse määramise teeb keeruliseks tema osiste sagedus ja polüseemsus.

Ühendverbide seas, mis said määratluse „ma ei tea“ üks või kaks korda, mingit seaduspära sageduse ja tähenduste vahel ei ole. Siinkohal tuleb aga arvesse võtta ka seda, et mõni vastaja võis ühendverbide määratlemisest lihtsalt ära tüdineda ning kindla numbrilise vastuse andmise asemel otsustas kompositsionaalsuse lihtsalt määramata jätta. Vaadeldes ühendverbidele antud hinnanguid ja nende standardhälvet, tundub, et selliseid ühendeid, mis on vastuse „ma ei tea“ saanud pigem kogemata, on kokku kaks ja nendeks on *peale langema* ja *alla kirjutama*. Mõlemad ühendverbid on kompositsionaalsuse määramisel saanud vaid kaks erinevat numbrilist kompositsionaalsuse määra ja ühe „ma ei tea“ vastuse. Tulemuste põhjal on need määratud pigem kompositsionaalseteks ühenditeks. Sellest tulenevalt on ka nende vastuste standardhälve madal – 0,53. Ülejäänud „ma ei tea“ vastuse saanud ühendverbide standardhälve on kõrgem ning inimeste poolt antud kompositsionaalsuse määrad väga erinevad, seega on vastus „ma ei tea“ tulnud tõenäoliselt õigustatult. Seda, kas mõni hindaja määras mitmele ühendverbile hinnangu „ma ei tea“ pole võimalik kasutatud andmestiku peale uurida.

Kokkuvõtlikult võib öelda, et raskesti hinnatavate ühendverbide sagedused varieeruvad ning ilmselget seost kompositsionaalsuse määramise keerukuse ja ühendverbi sageduse vahel ei leitud. Siiski võib analüüsi tulemusel öelda, et üheks põhjuseks, miks ühendverbi kompositsionaalsust on raske hinnata, on tema osiste kõrge sagedus ja mitmetähenduslikkus.

### 2.2.3. Polüseemsus ühendverbide kompositsionaalsuse mõjutajana

Nii varasemate uurimuste kui ka selles töös uuritud materjali põhjal ilmneb, et ühendverbide (ja teiste püsiühendite) kompositsionaalsuse hindamisel võib polüseemsus inimeste otsust kompositsionaalsuse hindamisel mõjutada. Selles peatükis võetaksegi vaatluse alla ühendverbide polüseemsuse mõju nende kompositsionaalsusele. Eraldi vaadeldakse andmestiku kõige sagedasemaid ühendverbe (vt tabel 3), kõrgeima ja madalaima kompositsionaalsusega ühendverbe (vt tabel 10), suure standardhälbega ühendverbe (vt tabel 7) ja neid ühendeid, mis said kompositsionaalsuse määratlemisel kõige rohkem „ma ei tea“ vastuseid (vt tabel 8).

Kõige sagedamini esinevate ühendverbide kompositsionaalsuse määr jääb vahemikku 2,36–4,17 (vt tabel 3). Skaalal ühest viieni jääb see määr keskmiseks ehk hindajad ei ole päris kindlad, kas nende ühendverbide tähendus tuleneb otseselt tema osiste tähendusest. Segadust võib tekitada nende ühendverbide polüseemsus ehk mitmetähenduslikkus, mille analüüsimiseks kasutati EKSSi ning vaadati, mitu tähendust neil viiel kõige sagedamini esineval ühendverbil on. Polüseemsuse mõju sagedaste ühendverbide keerukale kompositsionaalsuse määramisele on välja toonud ka teised sageduse mõju uurijad, nt Bott ja Schulte im Walde (2014). Tabelis 9 on välja toodud kümne sagedaima ühendverbi eri tähenduste arv EKSSi põhjal.

Tabel 9. **10 kõige sagedama ühendverbi polüseemsus**

| **Ühendverb** | **Sagedus** | **Tähendused EKSSis** |
|---|---|---|
| *vastu võtma* | 35 929 | 7 |
| *ette nägema* | 28 477 | 4 |
| *kinni pidama* | 19 757 | 6 |

| | | |
|---|---|---|
| *välja andma* | 19 743 | 5 |
| *kaasa tooma* | 17 743 | 2 |
| *kokku leppima* | 16 009 | 1 |
| *ette võtma* | 15 313 | 4 |
| *tagasi tulema* | 14 783 | 2 |
| *välja kuulutama* | 14 474 | 1 |
| *läbi viima* | 14 144 | 2 |

Tabelist 9 ilmneb väga selgelt, et sagedamini esinevatel ühendverbidel on EKSSi järgi mitu tähendust (EKSS). Näiteks on ühendverbil *vastu võtma* lausa seitse erinevat tähendust ning ühendil *kinni pidama* kuus erinevat tähendust. Teiste sõnapaaride hulgas jäävad silma *kokku leppima* ja *välja kuulutama*, millel on vaid üks tähendus, kuid kõigil ülejäänud ühendverbidel on neid vähemalt kaks. Seega võib järeldada, et sagedastel ühendverbidel on enamasti mitu tähendust. Vaadates nüüd uuesti nende ühendverbide kompositsionaalsust (vt tabel 3) võib järeldada, et sagedamini esinevate ühendverbide kompositsionaalsuse määr on keskmine seepärast, et ühendil on mitu tähendust, mis teeb ühese kompositsionaalsuse astme määramise keeruliseks.

Sellest et kõrge sagedusega ühendverbide kompositsionaalsust on raske hinnata nende polüseemsuse tõttu, võib omakorda järeldada, et kõik ühendverbid, mille kompositsionaalsuse määr oli väga kõrge ehk vahemikus 4,5–5 või vastupidi väga madal ehk vahemikus 1,9–2,3 (1,9 oli andmestiku kõige madalam kompositsionaalsus), on pigem ühetähenduslikud. Tabelis 10 on näha et suuremas osas ka nii on, sest neil ühendverbidel on maksimaalselt kolm eri tähendust ning needki tegelikult kõik küllaltki sarnased ehk on olemas üks prototüüpne tähendus. Näiteks ühendverbi *tagasi minema* kolm eri tähendust on *lähtekohta tagasi siirduma, tasemelt langema* ja *mingisse aega tagasi ulatuma*; ühendverbi *üles kloppima* kolm tähendust on aga *kloppides äratama, tagudes korda tegema (kõnekeeles)* ja *kohevaks kloppima* (EKSS).

Tabel 10. **Kuue kõrgeima ja madalaima kompositsionaalsuse määraga ühendverbi tähenduste arv EKSS-is**

| Ühendverb | Keskmine kompositsionaalsus | Tähendused EKSS-is |
|---|---|---|
| *eemale tõukama* | 4,67 | 2 |
| *ette jõudma* | 4,64 | 2 |
| *tagasi minema* | 4,58 | 3 |
| *järele kihutama* | 4,55 | 2 |
| *sisse kutsuma* | 4,54 | 1 |
| *kinni traageldama* | 4,50 | 1 |
| *välja nägema* | 1,90 | 2 |
| *vastu põrutama* | 1,92 | 2 |
| *üles kloppima* | 2,0 | 3 |
| *vahele kukkuma* | 2,18 | 1 |
| *taga kihutama* | 2,18 | 2 |
| *kinni maksma* | 2,29 | 2 |

Tabeli põhjal on näha, et ühendverbid, mis said hindajate poolt väga kõrge või väga madala kompositsionaalsuse hinnangu, on pigem ühe- või kahetähenduslikud. Inimesed olid nende ühendverbide määratlemisel küllaltki üksmeelel ning ühendverbi kompositsionaalsuses või mitte-kompositsionaalsuses oli vähe kahtlusi.

Kõrge standardhälbega ühendite seas (vt tabel 7) on väga selgelt näha, et neil on enamasti vaid üks tähendus. See tulemus on üllatav, sest kõrge standardhälve näitab, et inimesed ei olnud kompositsionaalsuse määramisel üksmeelel, millest võis eelnevate uurimuste (nt Reddy jt 2011) põhjal eeldada, et tegemist on polüseemsete ühenditega. Tuleb aga välja, et nii see ei ole. Töö autor usub, et ühendverbide standardhälve on kõrge seepärast, et iga testgrupis olija lähenes ühendite kompositsionaalsusele enda subjektiivsest vaatenurgast ning tabelis olevaid ühendverbe on küllaltki keeruline määratleda olenevalt sellest, kui põhjalikult neid analüüsida. Näiteks sõnapaar *vastu raiuma* on küllaltki

mittekompositsionaalne, sest füüsilise raiumisega siin tegemist ei ole ning mõeldakse pigem vastuhakku, vaidlemist või endale kindlaks jäämist. Samas on aga afiksaaladverb *vastu* väga otseselt seotud vastu hakkamise või vastu vaidlemisega, seega moodustub ühendverbi tähendus siiski selle osade tähendusest. Samuti on adverb *vastu* ise väga mitmetähenduslik ning EKSSi põhjal on sel lausa 12 erinevat tähendust. Kindlasti on see ka üheks asjaoluks, miks kompositsionaalsuse määramisega üksmeelel ei oldud. Samamoodi ühendverb *ümber mõtlema*, mille all mõeldakse otsuse/mõtte muutmist ning kus verb *mõtlema* on tähendusega tihedalt seotud, kuid afiksaaladverb *ümber* jällegi pigem mitte. Ka on adverbil *ümber* EKSSi põhjal 10 erinevat tähendust.

Ühendverbid, mis said kompositsionaalsuse määramisel kõige rohkem vastuseks „ma ei tea“, on EKSSi järgi pigem väikese tähenduste arvuga (vt tabel 8). Tabelis on küll ühendeid, mille tähenduste arv sõnaraamatu põhjal on kolm või neli, kuid need tähendused on omavahel kas väga sarnased või on mõni seletus kõnekeelne ning väga harva kasutusel. Sarnaselt eelnevas lõigus mainutule, usub töö autor, et ka selle rühma ühendverbide määramine oli olenemata ühest tähendusest keeruline seepärast, et ühendverb tundub esmapilgul mitte-kompositsionaalne, kuid pikema analüüsi tulemusel võib leida ühendi osade ja ühendverbi enda tähenduse vahel seoseid. Samuti on ka siin suureks mõjutajaks kindlasti adverbide polüseemsus, näiteks adverbil *üles* on EKSSi põhjal lausa 21 eri tähendusvarjundit. Seega kompositsionaalsuse määramine olenes sellest, kui põhjalikult inimesed ühendverbe analüüsisid ning milliseid seoseid nad tähenduste vahel lõid ehk näiteks millist adverbi tähendust analüüsiti.

Kindlasti olenes ühendverbide kompositsionaalsuse määramine näitelausetest, mis hinnatavate ühendverbide määramisel kuvati ning sellest kui hästi inimene ülesandest aru sai. Iga ühendverbi juures oli kolm näitelauset ning kui nendes oli ühel ja samal ühendverbil vastaja jaoks mitu erinevat tähendust, võiski kohe segadus tekkida. Madal sagedus korpuses võib tähendada ka seda, et inimesed ei puutu igapäevaelus nende ühendverbidega just tihti kokku ning seeläbi ka ei mõtle väga palju nende tähenduse moodustuse peale.

Suurimaks hinnangute mõjutajaks peab töö autor siiski inimeste subjektiivset analüüsivõimet ning ühendverbi osiste mitmetähenduslikkust. Edasiste uurimiste käigus peab sarnast küsitlust koostades jätma inimesele hindamiseks vaid ühe näitelause, milles tuleb kindlasti välja selle tähendus. See võib vähendada „ma ei tea“ vastuste arvu ning muuta hinnangud üksteisele sarnasemaks. Lisaks võiks analüüsida ka ühendverbide hindamist mõjutavaid faktoreid, kasutades psühholingvistilisi teste või uurida rohkem inimese igapäevase keelekasutuse kohta.

# Kokkuvõte

Töö käsitles sageduse mõju eesti keele ühendverbide tähenduse moodustumisel ning vaatles, kuidas sagedus mõjutab kompositsionaalsuse määra hindamist. Ühendverbide ja nende kompositsionaalsuse automaatne tuvastamine on oluline valdkond keeletehnoloogia ja leksikograafia jaoks. Ühe kompositsionaalsust mõjutava faktorina on teiste keele püsiühendite uurimisel vaadeldud ühendi sagedust, kuid seni polnud eesti keele peal seda veel põhjalikult uuritud.

Bakalaureusetöö eesmärk oli uurida kas ja mil määral mõjutab ühendverbide sagedus korpuses nende kompositsionaalsust. Töös kontrolliti väidet, et korpuses sagedamini esinevate ühendverbide kompositsionaalsuse määramine on keerulisem kui harva esinevate ühendverbide määramine, sest sagedased ühendverbid on tihti mitmetähenduslikud. Uuriti, kas ühendverbide sageduse ja kompositsionaalsuse vahel on seos ning kas polüseemsus mõjutab kompositsionaalsuse määramist.

Töö oli üles ehitatud kahest suuremast peatükist. Esimeses tutvustati eesti keele ühendverbi struktuuri, selle tähenduse moodustumist ja kompositsionaalsuse määramist. Toodi välja erinevad sellekohased uurimused Eestist ja mujalt maailmast. Teises peatükis anti ülevaade töö analüüsiosaks kasutatud materjalist ning analüüsi tulemustest. Välja toodi ühendverbide kompositsionaalsuse määramise, sageduse ja polüseemsuse vahelised seosed.

Ühendverbide analüüsimiseks kasutati varem koostatud kompositsionaalsuse andmestikku, mis koosnes 211 ühendverbist. Andmestik jagati kaheks ning eraldi uuriti ühendverbe, mille kompositsionaalsuse määra hindamist võis tinglikult pidada kergeks ülesandeks (157) ja neid, mille hindamist võis pidada keerukaks (54).

Töö ühe tulemusena leiti, et kasutatud andmestiku põhjal ei saa öelda, et ühendverbide ja selle komponentide sagedus mõjutab ühendverbide kompositsionaalsust. Samuti ei saa tulemuste põhjal väita ka vastupidist. Kõrge standardhälbe, kuid ühetähenduslike ühendverbide analüüs lubas oletada, et ühendverbide kompositsionaalsuse hindamise tegi keerukaks tema komponentide polüseemsus. Kõige keerulisemate ühendverbide ehk nende, mis said vähemalt kaks korda määratluseks „ma ei tea“ analüüsi tulemusena selgus, et ka need olid pigem ühetähenduslikud, kuid nende osised olid sagedased ja polüseemsed. Mitmetähenduslikud ühendverbi osad võisid inimestel ühendverbi analüüsides tekitada erinevaid tähendusseoseid ning seetõttu võis ka ühendi tähenduse analüüsimine keeruliseks osutuda.

Teise tulemusena leiti, et inimestel oli kompositsionaalsuse määramisel probleeme nii andmestiku kõige sagedasemate ühendverbidega kui ka küllaltki harva esinevate ühendverbidega (kõrge standardhälbega ja kõige rohkem „ma ei tea“ vastuseid saanud ühendverbid). Andmestiku kõige sagedasemad ühendverbid olid väga polüseemsed ning said kompositsionaalsuse skaalal ühest viieni keskmiseks väärtuseks 3. Sellest võib järeldada, et väga sagedaste ühendverbide kompositsionaalsuse ühene määramine oli keeruline just nende mitmetähenduslikkuse tõttu ning mitmetähenduslikele ühendverbidele ei saa ühte kokkuvõtvat hinnangut anda. Edasiste uurimuste tarbeks soovitas töö autor inimestele hindamiseks anda ühendverbi vaid ühe näitelausega, kus ühendil oleks üks kindel tähendus ning mis teeks kompositsionaalsuse määramise selle võrra kergemaks. See võib vähendada „ma ei tea“ vastuste arvu ning muuta hinnangud üksteisele sarnasemaks.

# Kirjandus


**Abedin jt = Abedin, Jaynal, Bipul Syam Purkayastha, Kh Raju Singha 2015**. Automated Multiword Expressions Detection in Bengali. – International Journal of Computer Science Engineering, 4(2), 56–61. http://www.ijcse.net/docs/IJCSE15-04-02-050.pdf. Vaadatud 25.05.2018.

**Aedmaa, Eleri 2015**. Statistilised meetodid ühendverbide tuvastamisel tekstikorpusest. – Eesti Rakenduslingvistika Ühingu aastaraamat, 11(0), 37–54. http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa11.03. Vaadatud 16.05.2018.

**Aedmaa, Eleri 2016**. Eesti keele ühendverbide kompositsionaalsuse määramine. – Eesti Rakenduslingvistika Ühingu Aastaraamat, 12(0), 5–23. http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa12.01. Vaadatud 20.05.2018.

**Aedmaa, Eleri 2017**. Exploring Compositionality of Estonian Particle Verbs. – Proceedings of the ESSLLI 2017 Student Session. 29th European Summer School in Logic, Language & Information July 17-28, 2017. Toulouse, France, 197–208. http://www2.sfs.nphil.uni-tuebingen.de/esslli-stus-2017/preproceedings_stus_2017.pdf#page=197. Vaadatud 25.05.2018.

**Bannard jt = Bannard, Colin, Timothy Baldwin, Alex Lascarides 2003**. A Statistical Approach to the Semantics of Verb-particles. – Proceedings of the ACL 2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment 18. Stroudsburg, PA, USA: Association for Computational Linguistics, 65–72. https://dl.acm.org/citation.cfm?doid=1119282.1119291. Vaadatud 18.04.2018.

**Bhattacharyya jt = Bhattacharyya, Pushpak, Sudha Bhingardive, Kevin Patel, Dhirendra Singh 2015**. Detection of Multiword Expressions for Hindi Language using Word Embeddings and WordNet-based Features. – Proceedings of the 12th International Conference on Natural Language Processing. NLP Association of India. Trivandrum, India, 295–302. https://www.cse.iitb.ac.in/~pb/papers/icon15-multi-word.pdf. Vaadatud 20.05.2018.

**Bott, Stefan, Sabine Schulte im Walde 2014**. Optimizing a Distributional Semantic Model for the Prediction of German Particle Verb Compositionality. – Proceedings of the 9th International Conference on Language Resources and Evaluation. European Language Resources Association (ELRA). Reykjavik, Iceland, 509–516. http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=457360A7EA66B8615B89494AAE7C0427?doi=10.1.1.696.8547&rep=rep1&type=pdf. Vaadatud 26.05.2018.

**Bullinaria, John A., Joseph P. Levy 2007**. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39(3), 510–526. https://link.springer.com/article/10.3758%2FBF03193020. Vaadatud 20.04.2018.

**Eesti keele koondkorpus 2015**. Tartu Ülikooli arvutilingvistika uurimisrühma koduleht. http://www.cl.ut.ee/korpused/segakorpus/. Vaadatud 07.04.2018.

**EKG II = Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvi Vare 1993**. Eesti keele grammatika II. Süntaks. Lisa: Kiri. Toim. Väino Klaus. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut, 18–22. http://dspace.ut.ee/handle/10062/29437. Vaadatud 22.04.2018.

**EKSS = Eesti keele seletav sõnaraamat 2009**. Toim. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks ja Piret Voll. http://www.eki.ee/dict/ekss/. Vaadatud 25.05.2018.

**Erelt, Mati 2013**. Eesti keele lauseõpetus: sissejuhatus. Öeldis. Tartu: Tartu Ülikooli eesti keele osakond, 19–64. http://dspace.ut.ee/bitstream/handle/10062/34069/erelt.indd.pdf?sequence=1. Vaadatud 20.04.2018.

**Goodkind, Adam, Andrew Rosenberg 2015**. Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production. – Proceedings of the 11th Workshop on Multiword Expressions. Denver, Colorado:

Association for Computational Linguistics, 87–95. http://www.aclweb.org/anthology/W15-0914. Vaadatud 12.05.2018.

**Kaalep, Heiki-Jaan, Kadri Muischnek 2002**. Püsiühendite leidmine teksti abil. Toim. Renate Pajusalu, Tiit Hennoste. Tähendusepüüdja: pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. TÜ üldkeeleteaduse õppetooli toimetised 3. Tartu: Tartu Ülikool, 172–184. http://dspace.ut.ee/handle/10062/58521. Vaadatud 25.05.2018.

**Kaalep, Heiki-Jaan, Kadri Muischnek 2009**. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. – Eesti Rakenduslingvistika Ühingu Aastaraamat, 5(0), 157–172. http://www.aclweb.org/anthology/W15-0914. Vaadatud 16.04.2018.

**Kallas, Jelena 2013**. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. Doktoritöö. Tallinn: Tallinna Ülikool. Humanitaarteaduste dissertatsioonid ISSN 1736-3624; 32, 10–185. https://erb.nlib.ee/?kid=29290314. Vaadatud 16.04.2018.

**Katz, Jerrold J., David Pitt 2000**. Compositional idioms. *Language* 76(2). Linguistic Society of America, 409-432. https://www.jstor.org/stable/417662?seq=1#page_scan_tab_contents. Vaadatud 20.05.2018.

**Katz, Graham, Eugenie Giesbrecht 2006**. Automatic Identification of Non-compositional Multi-word Expressions Using Latent Semantic Analysis. – Proceedings of the Workshop on Multiword Expressions: Identifying and Exploiting Underlying Properties. Stroudsburg, PA, USA: Association for Computational Linguistics, 12–19. http://dl.acm.org/citation.cfm?id=1613692.1613696. Vaadatud 21.05.2018.

**Kühner, Natalie, Sabine Schulte im Walde 2010**. Determining the Degree of Compositionality of German Particle Verbs by Clustering Approaches. – Semantic Approaches in Natural Language Processing. Proceedings of the Conference on Natural Language Processing 2010 (KONVENS). Universitätsverlag des Saarlandes. Saarland University Press , 47–56.

https://pdfs.semanticscholar.org/d086/d3838902a76faa66857d720170ec3fcf549a.pdf. Vaadatud 18.05.2016.

**McCarthy jt = McCarthy, Diana, John Carrol, Bill Keller 2003**. Detecting a Continuum of Compositionality in Phrasal Verbs. – Workshop on Multi-Word Expressions: Analysis, Acquisition and Treatment (ACL 2003), Sapporo, Japan, 73–80. https://dl.acm.org/citation.cfm?id=1119292 [aprill 2018]

**Piao jt = Piao, Scott S. L., Guangfan Sun, Paul Rayson, Qi Yuan 2006**. Automatic Extraction of Chinese Multiword Expressions with a Statis- tical Tool. – Proceedings of the Workshop on Multi-word-expressions in a multilingual context. Association for Computational Linguistics, 17–24. http://www.aclweb.org/anthology/W06-2403. Vaadatud 12.05.2018.

**Projekti eesmärgid ja tähtsus**. Eesti keele koondkorpus. Eesti keele keeletehnoloogiline tugi (2006–2010). https://www.keeletehnoloogia.ee/et/ekktt/ekktt-projektid/eesti-keele-koondkorpus. Vaadatud 30.03.2018.

**Reddy jt = Reddy, Siva, Suresh Manandhar, Diana McCarthy 2011**. An Empirical Study on Compositionality in Compound Nouns. – Proceedings of 5th International Joint Conference on Natural Language Processing. Chiang Mai, Thailand: Asian Federation of Natural Language Processing, 210–218. http://www.aclweb.org/anthology/I11-1024. Vaadatud 20.05.2018.

**Rätsep, Huno 1978**. Eesti keele lihtlausete tüübid. Eesti NSV Teaduste Akadeemia Emakeele Seltsi Toimetised 12. Tallinn: Valgus, 26–49. http://hunoratsep.ut.ee/wp-content/uploads/2012/11/ratsep_eesti_keele_ocr.pdf. Vaadatud 6.04.2018.

**Sag jt = Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger 2002**. Multiword Expressions: A Pain in the Neck for NLP. – Toim. Alexander Gelbukh. Computational Linguistics and Intelligent Text Processing 2276. Berlin, Heidelberg: Springer Berlin Heidelberg, 1–15. http://lingo.stanford.edu/pubs/WP-2001-03.pdf. Vaadatud 18.04.2018.

**Uiboaed, Kristel 2010**. Statistilised meetodid murdekorpuse ühendverbide tuvastamisel. – Eesti Rakenduslingvistika Ühingu Aastaraamat, 6(0), 307–326.

http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa6.19. Vaadatud 17.05.2018.

# The influence of the frequency on the compositionality of Estonian particle verbs. Summary

The present bachelor thesis investigates the influence of the frequency on the compositionality of Estonian particle verbs. In addition, it provides the analysis about the difficulties of the evaluation of the degree of compositionality.

Detecting multiword expressions (MWE) and their compositionality has been an important task for the natural language processing for years. Among other features, many authors have been studying the frequency of the expressions as a predictor of the compositionality of MWEs. This thesis is the first attempt to investigate the influence of the frequency on the compositionality of Estonian particle verbs.

The purpose of this thesis was to find out if and how the frequency of Estonian particle verbs influence their compositionality. The argument which this thesis was based on said that it was more difficult to determine the compositionality of particle verbs that are more frequent than it was for low frequency particle verbs, because high frequency particle verbs are usually ambiguous.

The first chapter of the thesis contained theoretical overview of the particle verbs, the automatic detection of them and their compositionality. It introduced the studies investigating the influence of the frequency on the compositionality of MWEs. The second chapter provided the statistical analysis of the correlation between the frequency and the compositionality. In addition, it contained the description of the most difficult particle verbs to evaluate and reports the influence of the polysemy on the evaluation of the degree of compositionality.

The statistical analysis revealed that the correlation between frequency and compositionality is low and statistically insignificant. Hence, the correlation between

these two features needs further investigation and based on this study, there is not enough evidence to confirm the absence of the correlation.

The analysis of the particle verbs that were difficult to evaluate showed that particle verbs with complicated compositionality were mostly unambiguous. The frequencies of those particle verbs were varying. The author believes that main reason why it was difficult for the evaluators to determine the degree of compositionality was that the components of the particle verbs were frequent and ambiguous. Hence, the ambiguity of the adverb and/or verb complicated peoples' ability to determine the degree of compositionality of particle verbs. As assumed, the most frequent particle verbs were very ambiguous which may be the reason that made the evaluation of their compositionality complicated. As a result, the average degree of compositionality of these particle verbs were often evaluated as 3 on the 5-point scale.

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina Monika Muru (sünnikuupäev: 9.07.1996)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Esinemissageduse mõju ühendverbide tähenduse moodustumisel“, mille juhendaja on Eleri Aedmaa.
    1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
    1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus 28. mail 2018