TARTU ÜLIKOOL
FILOSOOFIATEADUSKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Eleri Aedmaa

SÕNADEVAHELISE SEOSE TUGEVUSE MÕÕTMISE STATISTILISED MEETODID ÜHENDVERBIDE TUVASTAMISEL

Magistritöö

Juhendajad dotsent Kadri Muischnek ja teadur Kristel Uiboaed

TARTU 2014

SISUKORD

Tabelid

Joonised

Sissejuhatus

See magistritöö käsitleb sõnadevahelise seose tugevuse mõõtmise statistilisi meetodeid eesti keele ühendverbide automaatsel tuvastamisel tekstikorpusest.

Lingvistikas ei klassifitseerita sõnu mitte ainult nende tähenduste põhjal, vaid ka sellel alusel, milliste teiste sõnadega need koos esinevad (Church, Hanks 1990: 22). Sõnade sagedase koosesinemise põhjuseks võib olla lihtsalt nende endi suur sagedus tekstis, näiteks *see on*, *ja ka* jne, aga ka see, et nad moodustavad püsiva tavapärase sõnade ühendi keeles ehk püsiühendi (Kaalep, Muischnek 2009: 157–163). Arvutilingvistikas on *püsiühendi* mõiste kõrval rohkem kasutusel *kollokatsiooni* mõiste (Evert 2008: 3), mille näiteks Kaalep ja Muischnek (2002: 173) defineerivad sõnaühendiks, mida moodustavad sõnad esinevad tekstis koos sagedamini, kui võiks eeldada nende eraldi esinemise sageduse põhjal.

Keele automaattöötlemisel on kollokatsioonid problemaatiline nähtus, sest süntaktilise ja semantilise analüüsi jaoks on oluline mitmesõnalise üksuse või minimaalse semantilise üksuse äratundmine, mistõttu ei saa analüüsi aluseks võtta tühikutevahelist stringi, vaid kasutada tuleb teistsuguseid meetodeid (Kaalep, Muischnek 2009: 158). Üks kollokatsioonide tuvastamise võimalus on rakendada sõnadevaheline seose tugevuse mõõdikuid (*association measures*) (Wiechmann 2008: 257). Viimase 50 aasta jooksul on enim testitud sümmeetrilisi statistikud[1] (Gries 2013: 4), mis arvutavad igale korpusest leitud sõnapaarile ühe seose tugevuse väärtuse (*association score*), mis näitab kahe sõna vahelise statistilise seose suurust (Evert 2008: 5). Inimese meeles tekkivad sõnadevahelised seosed on oluline uurimisvaldkond psühholingvistikas (Church, Hanks 1990: 23) ning kuna üha enam korpuslingviste näevad keeleteadust kognitiivteaduste osana, siis lisaks sümmeetrilistele mõõdikutele on viimastel aastatel kollokatsioonide tuvastamisel arvestatud ka sõnadevahelise seose asümmeetrilisusega, mille tuvastamiseks rakendatakse asümmeetrilisi sõnadevahelise seose tugevuse mõõdikuid. Nende abil on

[1] Edaspidi kasutan sünonüümselt mõisteid sõnadevahelise seose tugevuse mõõtmise mõõdik ja statistik.

võimalik arvutada igale sõnapaarile kaks seose tugevuse väärtust, mis osutavad, milline sõna kollokaadis on rohkem mõjutatud teise sõna esinemisest. (Gries 2013: 5–13)

Minu töö eesmärk on välja selgitada, milline sõnadevahelise seose tugevuse mõõtmise meetod on kõige efektiivsem üht liiki eesti keele püsiühendite – ühendverbide – automaatseks tuvastamiseks tekstikorpusest, ning teada saada, kas statistikute tööd mõjutab tekstiliik ja/või korpuse suurus. Samuti on eesmärk testida asümmeetrilisi mõõdikuid, et uurida, kas ja kui otstarbekas on nende rakendamine eesti keele ühendverbide tuvastamisel tekstikorpusest.

Hüpotees on, et kõige parem meetod eesti keele ühendverbide tuvastamiseks on arvutilingvistikas sarnaste ülesannete lahendamisel palju kasutatud log-tõepära funktsioon (Evert 2004: 21), mis eesti murdetekstidega tehtud katsete põhjal andis parimaid tulemusi (Uiboaed 2010). Everti ja Krenni uurimuse (2001) tulemuste põhjal oletan, et statistikute paremusjärjestus on ühendverbide tuvastamisel erinevatest tekstiliikidest sarnane ning korpuse suurus ei mõjuta oluliselt mõõdikute tulemusi. Michelbacheri jt (2007; 2011) uurimuste tulemustele tuginedes eeldan, et asümmeetriliste statistikute rakendamine on tulemuslik ühendverbide tuvastamisel tekstikorpusest.

Töö jaguneb kolme suurema peatüki vahel. Esimene peatükk moodustab magistritöö teoreetilise osa ja keskendub püsiühendite tuvastamisele: annab ülevaate püsiühendi, kollokatsiooni ja ühendverbi mõistetest ning sellest, kuidas sõnadevahelist seost tuvastada ja mõõta. Viimane alapeatükk kirjeldab põgusalt varasemaid eesti keele püsiühendite automaatse tuvastamise katseid. Teine peatükk esitab ülevaate töö aluseks olevast andmestikust ning tulemuste hindamise meetoditest. Seejärel tutvustab uurimusse kaasatud sümmeetrilisi ja asümmeetrilisi mõõdikuid ja nende tööpõhimõtteid. Kolmas peatükk sisaldab sõnadevahelise seose tugevuse mõõtmiseks rakendatud statistiliste meetodite võrdluseid. Esimene neist vaatleb ja hindab väljavalitud statistikute tulemusi erinevaid tekstiliike sisaldavate korpuste põhjal, teine kirjeldab, millised on statistikute tulemused, kui ühendverbe tuvastatakse erineva suurusega korpustest. Kõrvutan ka sümmeetriliste ja asümmeetriliste mõõdikute tulemusi ning selgitan välja, kas asümmeetriliste statistikute kasutamine on põhjendatud eesti keele ühendverbide tuvastamisel. Viimases alapeatükis võrdlen teiste keelte põhjal tehtud sarnaste katsete tulemusi siinse uurimuse tulemustega.

1. Püsiühendite tuvastamine

Esimesena võtsid arvutilingvistikas termini *püsiühend* inglise keelse vaste *multiword expression* kasutusele Sag jt (2002: 2), kes defineerisid mõistet võrdlemisi üldiselt: püsiühend on mitmest sõnast koosnev ühend, mille tähendust ei saa järeldada selle komponentide tähenduste põhjal.

See peatükk annab ülevaate püsiühendi mõistest ja püsiühendite automaatse tuvastamise võimalustest. Pikemalt vaatlen arvutilingvistikas püsiühendiga võrdsustatud kollokatsiooni ja selle töö keskset ühendit ühendverbi. Seejärel kirjeldan sõnadevahelise seose tuvastamist ja tutvustan täpsemalt, milliseid andmeid on vaja statistiliste meetodite rakendamiseks. Tuginen põhiliselt Stefan Everti artiklile „Corpora and collocations" (2008), ühendverbi käsitlevas alapeatükis aga Huno Rätsepa monograafiale „Eesti keele lihtlause tüübid" (1978), „Eesti keele grammatikale" (1993), „Eesti keele käsiraamatule" (2007) ja Mati Erelti 2013. aastal ilmunud raamatule „Eesti keele lauseõpetus".

Peatüki lõpetab kokkuvõte varem tehtud eesti keele püsiühendite automaatse tuvastamise katsetest.

1.1. Püsiühendi mõiste

Püsiühend on sõnade ühend, mida mingi tähenduse väljendamiseks on tavaks koos kasutada. Need püsiühendid, mille kogutähendus ei tulene tema komponentide tähenduste summast, on idioomid (nt *silmas pidama*, *jalga laskma*, *südant puistama*). (Muischnek 2006: 12) Kui sõnad esinevad ühendis oma tavatähenduses, on tegu kollokatsiooniga, näiteks *silmi kissitama*. Verbikesksed püsiühendid jagunevad laiendi sõnaliigi alusel verbi ja (afiksaal)adverbi ühenditeks ehk ühendverbideks ning verbi ja noomeni(fraasi) või adpositsioonifraasi ühenditeks. (Kaalep, Muischnek 2009: 159) Selle töö seisukohalt olulise ühendverbi kirjelduse esitan peatükis 1.3.

Arvutilingvistikas on *püsiühendi* mõiste kasutusele võetud suhteliselt hiljuti, et määratleda paremini ebaselget *kollokatsiooni* mõistet (Evert 2008: 3). Järgnevalt esitangi ülevaate kollokatsiooni mõistest ja sellega seonduvast problemaatikast.

1.2. Kollokatsiooni mõiste

Kollokatsioon, üldiselt koosesinemine, on olnud paljude korpuslingvistiliste tööde keskmes aastakümneid (Gries 2013: 1). *Kollokatsioonil* on väga palju erinevaid seletusi, kuid selle defineerimine põhineb alati tähelepanekul, et teatud sõnadel on kalduvus esineda loomulikus keeles üksteise lähedal (Sinclair 1991: 71). Evert (2008: 2–3) toob välja mõiste *kollokatsioon* kolm võimalikku interpretatsiooni: fraseoloogilise, empiirilise ja selle, mis on kasutusel arvutilingvistikas, kus see on tihti üldistav nimetus igasuguste leksikaliseerunud semantiliste või süntaktiliste omadustega sõnakombinatsioonide kohta, mis vajavad masinloetavatelt sõnastikelt ja keeletöötlussüsteemidelt erilist kohtlemist. Sellisteks kombinatsioonideks on näiteks noomenifraasid (nt *Rootsi laud, lendav taldrik*), verbi ja tema seotud laiendi ühendid (nt *tähele panema*, *jalga laskma*) jne. Fraseoloogias on kollokatsioonideks poolkompositsioonilised ühendid, kus ühe komponendi kollokaatideks on teatud hulk sõnu, mida selle sõnaga koos kasutatakse, et edastada konkreetset mõtet. Empiirilise lähenemise jaoks on kollokatsioon sõnade tähendust ja kasutust iseloomustav vahend ehk lähtutakse sellest, et sõna kollokaatide põhjal on võimalik kirjeldada seda sõna ennast. Nende tähenduste eristamisel tuleb olla täpne ja segaduste vältimiseks peaks igas uurimuses kollokatsiooni mõiste määratlema just vastava töö kontekstis.

Selles töös käsitlen kollokatsiooni Sinclairi (1991: 71) definitsiooni järgi: kollokatsioon on kombinatsioon kahest sõnast, mis näitavad tendentsi esineda koos loomuliku keele tekstides. Selle töö uurimisobjekt – ühendverb – on seega püsiühendite hulka kuuluv kahest komponendist koosnev sõnaühend, mille tuvastamisel rakendan kollokatsioonide tuvastamisel kasutatavat sõnadevahelise seose tugevuse mõõtmist, mida kirjeldan täpsemalt peatükis 1.4. Enne seda annan ülevaate eesti keele ühendverbist ja selle käsitlusest varasemates töödes.

1.3. Ühendverb

Selles peatükis kirjeldan ühendverbi mõistet ja liigitust. Esimene alapeatükk sisaldab ülevaadet ühendverbi mõiste kujunemisest ning sellest, missugused on ühendverbi komponendid. Teine alapeatükk keskendub ühendverbide liigitusele, mis põhineb Huno Rätsepa (1978) käsitlusel.

1.3.1. Ühendverbi mõiste

Eesti keeles on tugev tendents analüütilisusele (EKK: 447) ja ilmselt osaliselt seetõttu ei läinud läbi Johannes Aaviku juba 1924. aastal tehtud ettepanek, mille järgi tuleb keelelise lühiduse saavutamiseks mitmesõnalised verbid asendada ühejuurelistega (Aavik 1974: 79), ja Elmar Muuk võttis artiklis „Verbide ja verbaalnoomenite kokkukirjutamisest“ (1938) kasutusele mõiste *ühendverb*. Ta toob välja kaks ühendverbi tunnust: ühendverbile on omane piltlik tähendus ja ühendverbis esineval verbil on ühendis selline tähendusvarjund, mida tal üksi esinedes ei ole (Muuk 1938: 8). Valgma ja Remmeli „Eesti keele grammatika“ (1968) järgi koosneb ühendverb kahest sõnast, mis kord kirjutatakse kokku ja kord lahku.

Rätsep, kes esitas esimesena põhjaliku ühendverbide käsitluse (1969, 1978), nimetab Valgma ja Remmeli definitsiooni ühendverbi ortograafilise kriteeriumi väljenduseks ja rõhutab, et ainult ortograafiline kriteerium pole ühendverbi defineerimiseks piisav (1978: 26). Nii toob ta lisaks välja veel morfoloogilise, süntaktilise ja semantilise kriteeriumi. Morfoloogilise kriteeriumi kohaselt koosnevad ühendverbid verbist ja abimäärsõnast, süntaktilise kriteeriumi järgi esinevad ühendverbid lauses ühe lauseliikmena ja semantilise kriteeriumi põhjal kirjutatakse verbidega kokku ainult adverbid, mis muudavad või täiendavad verbi tähendust või lisavad uue tähendusvarjundi. (Rätsep 1978: 26–27)

Pihlak (1985) kirjeldas eesti keele püsiühendeid (ühend- ja väljendverbe ning perifrastilisi verbe) aspekti analüütilise väljendamise vahendina ja võrdles eesti keele aspektilisi konstruktsioone vene keele omadega.

1993. aastal ilmunud „Eesti keele grammatika“ süntaksi-osa järgi (EKG II: 20) on ühendverbid perifrastilised verbid, mille sisuliseks tuumaks on verb, komplitseerivaks komponendiks aga adverb. EKG-l baseeruva „Eesti keele käsiraamatu“ (EKK: 446–447)

kohaselt võib ühendverbi adverbiks olla koha-, perfektiivsus-, seisundi- ja modaalmäärsõna.

Mati Erelt nimetab 2013. aastal ilmunud „Eesti keele lauseõpetuses“ kohamäärsõnu ka lokaalseteks määrsõnadeks, mis osutavad üldistatud kujul suhtelist suundumis-, paiknemis-, eemaldumis- või kulgemiskohta: *alla, asemele, eemale, ette, juurde, kaasa, kohale, kätte, külge, maha, otsa, pärale, peale, sisse, taha, vastu, üles, ümber; all, asemel, eemal, väljas; alt, asemelt, eemalt, väljast; läbi, mööda, ringi, üle* jne (Erelt 2013: 62).

Perfektiivsust väljendavad afiksaaladverbid märgivad tegevuse piiritletust või vähemalt piirivõimaluse olemasolu. Seesugused afiksaaladverbid on *läbi* (*lugema*), *maha* (*müüma*), *minema* (*minema*), *otsa* (*lõppema*), *tulema* (*tulema*), *täis* (*kritseldama*) jne. (Erelt 2013: 63–64) Kõige levinum perfektiivsusadverb on *ära* (Rätsep 1978: 31), mis moodustab hulga korrapäraseid ühendeid, nii selliseid, kus adverb täidab üksnes tegevuse piiritlemise ülesannet, nagu *ära jagama*, *ära mõistatama*, *ära ootama*, kui ka selliseid, kus aspektitähendus kaasneb suunatähendusega, nagu *ära kutsuma*, *ära viskama* (Erelt 2013: 64).

Seisundit väljendavad adverbid esinevad afiksaalsetena vaid siis, kui nad moodustavad koos verbiga uue tähendusliku terviku ja tingivad lausemalli, näiteks *kinni* (*nabima*), *lahti* (*saama*), *kokku* (*kukkuma*), *viltu* (*minema*/*kiskuma*) jms, modaalsust väljendavad afiksaaladverbid on *vaja* ja *tarvis* ühendverbides *vaja minema*/*olema*, *tarvis minema*/*olema* (Erelt 2013: 64).

1.3.2. Ühendverbi liigitus

Rätsepa (1978: 28) järgi jagunevad ühendverbid kahte rühma – ainukordseteks ja korrapärasteks. Selline jaotus on kõige selgem lokaalsete afiksaaladverbide juures (Erelt 2013: 62).

Ainukordsed ehk idiomaatilised ühendverbid koosnevad piiratud kombinatsioonivõimalustega osadest, mis moodustavad süntaktiliselt ja semantiliselt liigendamatu terviku. See tähendab, et 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: 21) Ainukordsete ühendverbide adverb kuulub verbaalsesse keskmesse, mis tähendab,

et adverb pole verbi seotud laiend ja selle kohta ei saa esitada eriküsimust (Rätsep 1978: 28).

Korrapärased ühendverbid ei ole erinevalt ainukordsetest ühendverbidest valmis sõnastikuüksused. Nad kujunevad mingi tähendusrühma verbide suhteliselt regulaarsel kombineerumisel kindlasse rühma kuuluvate afiksaaladverbidega: *alla* / *eemale* / *juurde* / *kohale* / *ligi* / *pärale* / *tagasi* … + *minema* / *jooksma* / *astuma* / *kihutama* / *sõitma*. (EKG II: 21) Korrapäraste ühendverbide adverbid on verbi seotud laiendid, mille kohta on võimalik esitada ka eriküsimus (Rätsep 1978: 29). Mõlemad ühendi osised säilitavad tähendusliku iseseisvuse, kuid vaatamata sellele moodustavad ka korrapärased ühendverbid süntaktiliselt lahutamatu terviku. Lausemalli määrab finiitverb koos afiksaaladverbiga. (EKG II: 21)

See töö ei erista ainukordseid ja korrapäraseid ühendverbe – kõikide ühendverbide tuvastamise aluseks võtan sõnadevahelise seose, ning tulemusi erinevat liiki ühendverbide seas ei kõrvuta. Tulemuste võrdlemine ühendverbide liikide seas võiks olla selle töö edasiarenduse osa.

Järgmises peatükis kirjeldan sõnadevahelist seost ning selle tugevuse mõõtmist.

1.4. Sõnadevaheline seos ja selle tugevuse mõõtmine

Kui kollokatsiooni definitsioon põhineb sõnade koosesinemisel, siis järgmisena kirjeldan täpsemalt, mida selle all mõtlen ning määratlen eelnevate uurimuste põhjal, mida tähendab seos sõnade vahel ja kuidas selle tugevust mõõta. Täpsemalt teen ülevaate lähenemisviisidest sõnadevahelise seose mõõtmiseks, millest võib lähtuda kollokatsioonide kandidaatpaaride moodustamisel, ning sellest, millist andmestikku sõnadevahelise seose tugevuse mõõtmiseks vaja on.

1.4.1. Lähenemisviisid kandidaatpaaride moodustamiseks

Evert (2008: 12–16) toob välja kolm sõnadevahelise seose mõõtmise lähenemisviisi, millest võib lähtuda püsiühendite kandidaatpaaride moodustamisel.

Kindlas naabruses koosesinemine (*surface cooccurrence*) on sõnadevahelise tugevuse mõõtmisel kõige levinum lähenemisviis. Selle järgi loetakse sõnu koosesinevateks, kui nad esinevad koos mingis kindlas kauguses või samas aknas (*collocational span*), mida mõõdetakse kahe sõna vahel olevate üksuste kaudu (näiteks

sõnad, tähemärgid jne). Seejuures on uurijal vaja teha mitmeid otsuseid. Esiteks tuleb määrata akna suurus, mis tavaliselt jääb kolme ja viie sõna vahele. Lisaks tuleb otsustada, kas arvestatakse kõiki üksuseid (kaasa arvatud kirjavahemärke ja numbreid) või ainult sõnu. Samuti tuleb otsustada, kuidas käituda mitmesõnaliste ühenditega ja kas koosesinemine võib ületada lausepiire. (Evert 2008: 12)

Akent tähistatakse üldiselt vormis (Lk, Rn), kus k tähistab üksuste arvu, mis jäävad uuritavast sõnast vasakule ja n tähistab üksuste arvu, mis jäävad paremale. Seega võib aken olla nii sümmeetriline, mis tähendab, et uuritavast sõnast jääb nii vasakule kui ka paremale sama arv üksuseid (näiteks L4, R4), kui ka asümmeetriline, mis tähendab, et uuritava sõna vasakul ja paremal pool on erinev arv üksuseid (näiteks L2, R4).

Tekstuaalse koosesinemise (*textual cooccurence*) korral moodustavad kaks sõna potentsiaalse püsiühendi, kui nad esinevad ühes ja samas tekstiüksuses, tüüpiliselt lauses või lausungis (Evert 2008: 13). Kui vaadeldavaks tekstiüksuseks on lause, siis kõik lauses esinevad sõnad võivad moodustada kollokatsiooni teise lauses esineva sõnaga. Lauses *See raamat, mida ma eile otsisin, on vanaema juures* moodustavad potentsiaalse püsiühendi näiteks sõnad *see raamat, see mida, raamat mida, raamat eile* jne. Kui vaadeldavaks tekstiüksuseks on sama lause puhul osalause, siis moodustavad potentsiaalse püsiühendi näiteks sõnad *see raamat, mida ma, ma eile, vanaema juures* aga mitte sõnad *eile vanaema* või *raamat otsisin*.

Erinevalt kindlas naabruses koosesinemise lähenemisest, mille puuduseks peetakse akna suuruse valiku suvalisust, arvestab tekstuaalne lähenemine rohkem vaba sõnajärjega keeltega, kus omavahel seotud sõnad võivad paikneda teineteisest kaugel. Järelikult sobib siin aknaks lause või osalause. Lisaks võimaldab tekstuaalne lähenemine tuvastada nõrgemalt seotud sõnaühendeid. (Evert 2008: 13–14)

Süntaktiline koosesinemine (*syntactic cooccurance*) on kõige suuremate piirangutega, selles loetakse sõnad koosesinevateks vaid juhul, kui nende vahel on kindel süntaktiline seos. Selline suhe on näiteks verbi ja selle objekti või subjekti vahel. Püsiühendeid vaadeldakse eraldi lähtuvalt süntaktilise seose tüübist. Süntaktiline lähenemine on sobilik eelkõige juhul, kui kollokaadid asuvad üksteisest kaugel: erinevalt kindlas naabruses koosesinemise lähenemisest ei rakendata mingit üksustevahelise kauguse piirangut ja tulemus ei sisalda nii palju n-ö müra, kui tekstuaalse lähenemise

rakendamisel tekkida võib. Süntaktilist lähenemist kasutatakse sageli püsiühendite tuvastamisel, sest paljud leksikaliseerunud ühendid esinevad koos kindlas süntaktilises seoses, näiteks adverbi ja verbi ühend võib koos moodustada öeldise (nt *ülal pidama*, *läbi põlema*). (Evert 2008: 14–15)

Kokkuvõtlikult võib öelda, et kindlas naabruses koosesinemise lähenemine ei eelda märgendatud korpust ja selle rakendamine on osutunud tulemuslikuks korpuslingvistikas ja leksikograafias (Evert 2008: 16). Märgendamine on töö- ja ajamahukas (Garside jt 1997: 5) ning seetõttu on vabalt kättesaadavaid märgendatud korpuseid vähe ja need on väikesed, märgendamata korpused on aga suured ja vabalt kättesaadavad, mistõttu kindlas naabruses koosesinemine on leidnud laia kasutust. Tekstuaalset lähenemist, mille rakendamiseks on vajalik lausepiiridega märgendatud korpus, on mõistlik kasutada erinevat tüüpi juhuslike väljendite välistamiseks. Süntaktiline lähenemine on otstarbekas kindla konstruktsiooniga püsiühendite leidmisel, kuid eeldab süntaktiliselt märgendatud korpust. (Evert 2008: 16)

1.4.2. Sõnadevahelise seose tugevuse mõõtmine

Kui sõnadevaheline seos on defineeritud vaid nende sagedase koosesinemise põhjal, siis võib öelda, et igasuguste sõnade paar, mis esineb koos vähemalt kaks korda, on potentsiaalne kollokatsioon. Seepärast on tavaline rakendada kõrgemat sagedusläve (*frequency threshold*), selleks võib olla näiteks 3, 5 või isegi 10 koosesinemist. (Evert 2008: 5) Selles töös pole ma kõrgemat sagedusläve rakendanud ja kandidaatpaaride hulgas on ka vaid korra korpuses esinenud ühendid, sest nende seas on õigeid ühendverbe ning osa kasulikust informatsioonist oleks kaduma läinud.

Ainult kordumine ei ole piisav alus, et sõnadevahelist seost tugevaks pidada. Sõnadevahelise seose tugevuse statistikute rakendamine on vajalik, sest need aitavad määrata, kas tegemist on „õige kollokatsiooniga“ ja kas kollokatiivne seos on „tugev“ või „nõrk“. (Evert 2008: 5) Seose tugevuse mõõtmiseks koosesinevate sõnade vahel saab kasutada matemaatilisi seose tugevuse mõõdikuid. Nende kasutamise tulemuseks on iga sõnapaari jaoks seose tugevuse skoor, kus tavaliselt kõrge skoor viitab tugevale ja madal skoor nõrgale seosele. Seose tugevuse skoore kasutatakse, et välja valida „õiged kollokatsioonid“ rakendades sageduse lävendit või järjestades sõnapaarid skoori järgi

kahanevalt (seega leiab tugevamad kollokatsioonid nimekirja eesotsast). (Pecina, Schlesinger 2006: 652)

Selleks et sõnadevahelise seose tugevuse mõõtmise statistikuid rakendada, tuleb esmalt defineerida, mida mõeldakse sõnade koosesinemise all. Nii tuleb otsustada, kas otsitakse vaid n-ö tõelisi kollokatsioone või tahetakse saada ülevaadet kõikidest sõnapaaridest, asetades need sõnadevahelise tugevuse järgi mingile skaalale, eristamata kollokatsioone ja mitte-kollokatsioone. Esimese lähenemise korral peab uurija ise määrama seose tugevuse piirväärtuse, millest ülespoole jäävad sõnapaarid on n-ö tõelised püsiühendid. Teine otsus puudutab kollokatsioonide grupeerimist: kas otsitakse kõige tugevamini seotud sõnade paare või huvitutakse mingi sõna kindlast kontekstist ehk uuritakse missuguste sõnadega vaadeldav sõna kõige rohkem koos esineb. Kaks otsust on küll üksteisest sõltumatud, kuid sõna konteksti otsimine kombineeritakse tihti sõnapaaride mingisugusele skaalale asetamisega. Kui on tehtud vajalikud otsused, siis saab erinevate sõnadevahelise seose tugevuse mõõtmise meetodite abil leida sõnadevahelise seose tugevuse väärtused, mille interpretatsioon on selge ja lihtne: suuremad väärtused osutavad sõnadevahelisele tugevamale seosele, nõrgemad sellele, et sõnapaari kuuluvad sõnad pigem väldivad koosesinemist. (Evert 2008: 6)

Kõige lihtsam meetod kollokatsioonide tuvastamiseks tekstikorpusest on nende kokkulugemine ehk leida sõnade koosesinemise arv. See lähenemine töötab eriti hästi kindla konstruktsiooniga kollokatsioonide peal (nt *jalga laskma*, *tantsu lööma)*. Teistsuguste konstruktsioonide jaoks pole sõnade koosesinemise sagedus sõnadevahelise seose tugevuse väärtusena piisav. (Manning, Schütze 1999: 153–157) Näiteks kui moodustada korpuse sõnapaaride sagedusloend, siis ilmselt oleks selles üsna kõrgel kohal sõnapaar *ja ei*, mis tegelikult püsiühend ei ole. Kuna mõlemad sõnad on korpuses sagedased, siis on ka nende koosesinemine sage. Järelikult tuleb kasutada keerulisemaid meetodeid ja lisaks sõnapaari sagedusele arvesse võtta mõlema sõnapaari kuuluva sõna sagedused ehk marginaal- ehk ääresagedused, arvestada tuleb ka valimimahtu ehk korpuse suurust, kust püsiühendid leitakse (Evert 2008: 17).

Nii moodustubki sõnadevahelise seose tugevuse mõõtmiseks kasutatavate statistikute jaoks vajalik andmestik: O on sõnade koosesinemise sagedus valimis, f_1 ja f_2 on vastavalt esimese ja teise sõnapaari kuuluva sõna marginaalsagedus ja N on

valimimaht (Evert 2004: 36). Lisaks nõuavad statistilised meetodid ka teoreetilist sagedust (*expected frequency*) E, mis osutab sõnade koosesinemise teoreetilisele tõenäosusele. Selle arvutamiseks tuleb sõnade marginaalsageduste korrutis jagada valimi suurusega: $E = f_1 * f_2 / N$. (Evert 2008: 18)

Suur hulk statistilisi meetodeid kasutab kahemõõtmelist sagedustabelit (*contingency table*), mis arvestab marginaalsagedustega. Tabel 1 esitab näite kahemõõtmelisest sagedustabelist koos tehetega, mis illustreerivad vajalike väärtuste leidmist.

Tabel 1. **Kahemõõtmeline sagedustabel** (Evert 2008).

O_{11}	**O**	**f_1 – O**	O_{12}
O_{21}	**f_2 – O**	**N – f_1 – f_2 + O**	O_{22}

1.5. Varasemad eesti keele püsiühendite automaatse tuvastamise katsed

Üksikuid katseid eesti keele kollokatsioonide tuvastamiseks on tehtud ka varem. Kaalep ja Muischnek (2002) püüdsid tuvastada eestikeelsest tekstikorpusest püsiühendeid, täpsemalt ühend- ja väljendverbe kombineerides lingvistilisi ja statistilisi meetodeid. Eesmärgi saavutamiseks kasutasid autorid sõnadevahelise seose tugevuse hindamisel ühise oodatavuse määra koos GenLocalMax algoritmiga, mis realiseeriti eesti keele verbikesksete ühendite leidmiseks tarkvara SENVA abil. Katse käigus üritasid autorid vastata küsimustele 1) Kas tekstides leidub püsiühendeid, mida sõnaraamatutes ei esitata? ja 2) Kui lai on erinevate püsiühendite kasutusala? Selleks et kontrollida programmi tööd ja hinnata püsiühendite andmebaasi, võrreldi programmi abil saadud fraasiverbe püsiühendite andmebaasiga[2]. Selgus, et tekstikorpusest fraasiverbide leidmine ei ole triviaalne ja väljund vajab käsitsi toimetamist. Tulemustest selgus, et probleemidele vaatamata sobib SENVA hästi vaba sõnajärje ja keeruka morfoloogiaga eesti keelele rakendamiseks.

Uiboaed (2010) rakendas statistikuid Eesti murrete korpuse kaheliikmeliste ühendverbide automaatselt tuvastamisel ning katsetas kolme murderühma peal eraldi

[2] http://www.cl.ut.ee/ee/ressursid/pysiyhendid.html

nelja statistikut: t-skoori, vastastikuse informatsiooni väärtust (MI), hii-ruut statistikut ja log-tõepära funktsiooni. Uurimuse eesmärk oli eelkõige tutvustada statistiliste meetodite tööpõhimõtteid ning rakendada neid murdekorpuse morfoloogiliselt märgendatud ja automaatselt osalausestatud tekstide peal. Töö tulemusena selgus, et omavahel sarnasemaid tulemusi andsid hii-ruut-statistik ja MI ning t-skoor ja log-tõepära funktsioon. Ilmnes ka, et hii-ruut-statistik ja MI eelistavad selgelt madala esinemissagedusega ühendeid, mille komponentide esinemissagedused korpuses on samuti madalad. Uurimus kinnitas ka, et ühtegi mõõdikut ei saa pidada teistest ühemõtteliselt paremaks ning erinevat tüüpi statistikud sobivad erinevat tüüpi ülesannete lahendamiseks. Murdematerjali peal töötas küll kõige paremini log-tõepära funktsioon, ent autor ei soovita valida ühte kindlat mõõdikut kogu materjali jaoks.

Lisaks loodi 2010. aastal riikliku programmi „Eesti keele keeletehnoloogiline tugi 2010“ projekti „Eesti keele koondkorpuse esituse ja kasutusvõimaluste arendamine“ raames automaatne veebis kasutatav kollokatsioonide leidja Tasakaalus korpusest[3]. Rakendatud on kolme statistikut: log-tõepära funktsiooni, vastastikuse informatsiooni väärtust ja minimaalset tundlikkust. Lisaks on võimalik reastada sõnapaare esinemissageduse põhjal.

[3] https://korpused.keeleressursid.ee/clc/

2. Sõnadevahelise seose tugevuse mõõtmine

Selles peatükis esitan esmalt ülevaate töö uurimismaterjalist ja kogumise viisist, seejärel tutvustan tulemuste hindamise meetodeid. Teises alapeatükis kirjeldan töösse valitud sõnadevahelise seose tugevuse mõõtmise statistikuid ja nende tööpõhimõtteid.

2.1. Materjal ja meetodid

Peatükk annab ülevaate sõnadevahelise seose tugevuse mõõtmiseks kasutatud materjalist ja selle kogumise viisist. Seejärel tutvustan, millised on selle töö tulemuste hindamiseks kasutatud meetodid.

2.1.1. Materjal

Selle töö uurimismaterjal on Eesti keele koondkorpuse[4] tekstid – kokku 180 miljonit sõna. Erinevate tekstiliikide võrdlemine põhineb koondkorpuse allosa Tasakaalus korpuse[5] tekstidel, mis sisaldavad 15 miljonit sõna. Iga tekstiliiki – aja-, ilu- ja teaduskirjandust – on selles korpuses võrdselt ehk 5 miljonit sõna. Korpuse mahu suurenemise mõju vaatlus põhineb 170 miljonist sõnast koosneva ajakirjandustekstide valimil. Materjal oli eelnevalt t3mestaga (Kaalep 1998; Kaalep, Vaino 1998) morfoloogiliselt analüüsitud, ühestatud ning (osa)lausestatud (Kaalep, Muischnek 2012).

Püsiühendite tekstist tuvastamist alustasin ühendikandidaatide (kõikvõimalike potentsiaalsete ühendverbide) moodustamisega. Ühendverbide kandidaatpaaride moodustamisel lähtusin tekstuaalsest koosesinemisest (vt ptk 1.4.1.2) – aknaks oli siin osalause, sest ühendverbi moodustavad sõnad saavad esineda vaid samas osalauses. Teksti sõnajärge kandidaatpaaride moodustamisel ma ei arvestanud ning kuna eesmärk on tuvastada ühendverbe, siis iga osalause sees genereeriti adverbi ja verbi paaride kõikvõimalikud kombinatsioonid, mis moodustavad kandidaatpaaride loetelu. Selle nimekirja peal rakendasin stopp-sõnade loendit ehk eemaldasin ühendid, milles esinev

[4] http://www.cl.ut.ee/korpused/segakorpus/index.php
[5] http://www.cl.ut.ee/korpused/grammatikakorpus/index.php.

adverb reeglina ühendverbi koosseisus ei esine (nt *ikka*, *jälle*). Stopp-sõnade loend põhineb „Eesti keele seletava sõnaraamatu“ (EKSS) ühendverbide loendil[6], mis ühtlasi on selles töös ka kuldstandardiks ehk õigete ühendverbide loeteluks, millega saadud tulemusi edaspidi võrdlen. EKSS-i ühendverbide loendist eemaldasin ühendverbid, mis ei ole selles töös kasutatud korpuse märgendamise seisukohalt võimalikud. Sellisteks ühenditeks on sõnavormiga *kätte* moodustatud ühendverbid, nt *kätte jõudma*, *kätte maksma*, *kätte saama*, aga ka sellised ühendverbid nagu *minema kihutama*/*minema*/*viskama* ja *tulema tulema*. Nimetatud komponendid ei saa selles töös kasutatud korpuse märgenduses mitte kunagi adverbi märgendit ja seega on nendega võimatu moodustada kasulikke kandidaatpaare. N-ö müra vähendamiseks eemaldasin 1757 ühendverbi sisaldavast algsest EKSS-i loendist 20 mittekasulikku ühendit ning nii sisaldab selles töös kasutatav loend 1737 ühendverbi. Kandidaatpaaride loendi tekitamisel võtsin arvesse vaid EKSS-i ühendite adverbilise komponendi: kui adverbi EKSS-i nimistus ei esinenud, viskasin adverbi sisaldava ühendi kandidaatpaaride loetelust välja.

2.1.2. Täpsus ja saagis

Seda, kui tõhus on valitud ühendverbide leidmise meetod, saab hinnata täpsuse ja saagise arvutamisega. Täpsus kirjeldab leitud õigete ühendverbide suhet kõigi leitud ühendite hulgaga ja näitab, kui suur osa leitud ühenditest on õiged ühendverbid. Täpsus jääb 0% ja 100% vahele: 0% tähendab, et ükski leitud ühenditest pole õige ehk EKSS-i ühendverb, 100% näitab, et kõik leitud ühendid on EKSS-i ühendverbid. Saagis väljendab kõigi meetodiga tuvastatud õigete ühendverbide suhet kõigi võimalike õigete ühendverbidega (siin töös EKSS-i ühendverbidega) ning kirjeldab, kui suurt osa õigetest ühendverbidest õnnestus meetodiga andmestikust leida. Saagis jääb samuti 0% ja 100% vahele ning 0% tähendab, et ei leitud ühtegi õiget ehk EKSS-i ühendverbi, 100% aga seda, et on leitud kõik EKSS-is olevad ühendverbid.

Tabel 2 esitab kogu siinse töö materjali andmed: materjal koosneb 180 miljonist sõnast ja 25 912 251 osalausest, materjalist genereeritakse 71 332 kandidaatpaari ning

[6] Loend on saadud Eesti Keele Instituudist. Aitäh Jelena Kallasele abi eest! EKSS-i ühendverbide täielik loend on kättesaadav aadressil http://kodu.ut.ee/~eleriaed/magistrit88_failid/EKSS_loend.pdf

õigeid ühendverbe on nende seas 1684. Kokku on õigeid ühendverbe võimalik tuvastada 1737.

Tabel 2. **Ülevaade töös kasutatud materjalist.**

kogu töös kasutatud materjal	
sõnu	180 000 000
osalauseid	25 912 251
kandidaatpaare	71 332
tuvastatud õigeid ühendverbe	1684
õigeid ühendverbe kuldstandardis	1737
täpsus	(1684/71332)*100 ≈ 2,4%
saagis	(1684/1737)*100 ≈ 96,9%

Sageduse täpsus kogu materjalist ühendverbide tuvastamisel näitab, et 2,4% leitud ühenditest on EKSS-i ühendverbid. Saagise väärtus väljendab, et tuvastati 96,9% kõigist EKSS-i ühendverbidest ehk tuvastamata jäi 53 EKSS-i nimistusse kuuluvat ühendverbi. Sagedusloendi kasutamine tagab suure hulga õigete ühendverbide leidmise, kuid tuvastatud ühendite koguhulgast moodustavad õiged ühendverbid väikese osa.

2.1.3. Täpsuse kõverad

Sümmeetriliste statistikute tulemuslikkuse hindamiseks on Krenn ja Evert (2001) kasutanud täpsuse kõveraid (*precision curves*). Selleks arvutatakse väljavalitud mõõdikutega iga sõnapaari jaoks statistilise seose tugevuse väärtused ning saadud väärtuste järgi reastatakse sõnapaarid kahanevalt ümber. Uue loendi põhjal loetakse koosesinevateks vaid teatav hulk ühendeid sagedusloendi esimesest osast (selle väärtuse, millest alates ühendeid enam koosesinevateks ei loeta, peab uurija määrama ise). Statistikute väärtuste põhjal tehtud pingeridu kõrvutatakse õigete kollokatsioonide loendiga ning igale võimalikule kandidaatpaaride arvule (*n-best*) leitakse statistiku põhjal arvutatud tulemuste järgi selle täpsus. (Krenn, Evert 2001: 39–41)

Näiteks võetakse 500 esimest sõnapaari, mis on saanud statistiku t-skoor kõrgemad väärtused. Seejärel arvutatakse välja täpsus ehk mitu protsenti nendest 500 ühendist kuulub õigete kollokatsioonide loendisse (siinses töös EKSS-i ühendverbide loetellu). Sama saab teha kõikide valitud meetoditega ning erineva arvu kandidaatpaaridega. Selleks et määratleda uuritavate meetodite tulemuslikkust püstitatud ülesande lahendamisel, arvutatakse välja algtaseme täpsus (*baseline precision*) ehk teoreetiline

täpsus. Algtaseme täpsus arvutatakse korpusest leitud õigete ühendverbide arvu ja genereeritud kandidaatpaaride arvu põhjal. Seejärel kantakse saadud tulemused täpsuse kõveratena joonisele, kus iga kõver märgib erinevat mõõdiku täpsust.

Kui statistiku täpsuse kõver asub valdavalt teoreetilist täpsust märkivast joonest kõrgemal, siis võib mõõdikut pidada ülesande lahendamisel tulemuslikuks ja statistikut tasub kasutada ülesande edukaks sooritamiseks. Kui statistiku täpsuse kõver asub valdavalt allpool teoreetilist täpsust märkivast joonest, siis ei ole mõõdiku rakendamine ülesande lahendamiseks mõttekas. (Krenn, Evert 2001: 39–41)

Lisaks algtaseme täpsusele võib statistikuid võrrelda ka teiste näitajatega. Kuna mitmed tööd (nt Krenn, Evert 2001; Wermter, Hahn 2006) on tõestanud ka lihtsa sageduse tõhusust püsiühendite tuvastamise katsetes, siis on siinses töös uuritud, millised on lihtsa sagedusloendi tulemused võrreldes teiste meetoditega.

2.1.4. Sümmeetriliste ja asümmeetriliste mõõdikute võrdlemisest

Sümmeetriliste ja asümmeetriliste mõõdikute (vt lähemalt ptk 2.2) tulemuste võrdlus põhineb selles töös statistikute tuvastatud õigete ühendverbide loetelude kõrvutamisel. Kuna eesmärk on välja selgitada, kas ja kuidas erinevad asümmeetriliste statistikute tulemused sümmeetriliste statistikute omadest, siis selgitan välja, kas asümmeetrilised mõõdikud suudavad võrdse arvu kandidaatpaaride seast tuvastada rohkem ja/või teistsuguseid ühendverbe kui sümmeetrilised statistikud.

Tulemuste võrdlemiseks moodustan kaks loetelu: sümmeetriliste statistikute tuvastatud õigete ühendverbide loetelu ja asümmeetriliste statistikute tuvastatud õigete ühendverbide loetelu. Need loetelud koostan 50 kandidaatpaari hulgast ja aluseks võtan statistikute väärtuste kahaneva pingerea. Näiteks tuvastan 50 kõrgeima t-skoori väärtuse saanud kandidaatpaari seast õiged ühendverbid, mille liidan teiste sümmeetriliste mõõdikute 50 kõrgeima väärtuse saanud kandidaatpaaride hulka kuuluvate õigete ühendverbidega. Seejärel moodustan loetelu sümmeetriliste mõõdikute tuvastatud õigetest ühendverbidest, kus iga ühendit esineb vaid korra. Sama teen asümmeetriliste statistikutega. Saadud loendite võrdluse käigus selgub, kas ja kui palju on selliseid ühendeid, mida suudavad tuvastada ainult üht liiki statistikud.

Lisaks võrdlen Griesi (2013: 7–8) eeskujul sümmeetriliste ja asümmeetriliste statistikute tulemusi asümmeetrilise ΔP tulemuste abil. Kui Gries võttis tulemuste

võrdlemisel aluseks need ühendid, mille kahe ΔP väärtuse vahe $\Delta P(word_2|word_1)$ – $\Delta P(word_1|word_2)$ on suur (≥0,5 või ≤0,5), siis mina eraldan ühendid, mille ΔP kahe väärtuse vahe on >0 või <0, sest neid ühendeid võib pidada asümmeetrilisteks. Nende ühendite seast eraldan omakorda õiged ühendverbid ning õigete ühendverbide nimekirja võrdlen sümmeetriliste statistikute sama arvu kandidaatpaaride seast leitud õigete ühendverbide loendiga.

Näiteks kui neid ühendeid, mille ΔP väärtused erinevad, on kokku 50, siis eraldan iga sümmeetrilise statistiku 50 kõrgeima väärtuse saanud kandidaatpaari hulgast õiged ühendverbid ja liidan kõikide sümmeetriliste mõõdikute 50 kõrgeima väärtusega ühendi hulgast tuvastatud õiged ühendverbid selliseks loeteluks, kus kõiki tuvastatud ühendeid esineb korra.

Selleks et välja selgitada, kas asümmeetrilise ΔP tulemused on erinevad sümmeetriliste statistikute tulemustest, võrdlen kahte õigete ühendverbide loetelu ja vaatlen, kas ΔP tuvastatud ühendverbide seas on selliseid ühendeid, mida sümmeetriliste mõõdikute loetelus pole ehk kas ΔP üksi suudab tuvastada selliseid ühendeid, mida sümmeetrilised kokku ei suuda.

Enne tulemuste esitamist tutvustan aga neid sõnadevahelise seose tugevuse mõõdikuid, mille tulemusi vaatlema ja hindama hakkan.

2.2. Sõnadevahelise seose tugevuse mõõtmise meetodid

Sõnadevahelise seose tugevuse mõõdikud on statistilised valemid, mille abil saab arvutada sõnadevahelise seose statistilise tugevuse (Evert 2008: 5). Sõnadevahelise seose tugevuse mõõdikuid on palju ja nende rakendusalad lingvistikas on erinevad, kuid kõige rohkem on neid kasutatud kollokatiivsust puudutavates uurimustes. Kollokatsioonide tuvastamine statistikute abil on osutunud tulemuslikuks arvukates kontekstides: näiteks lähisünonüümide eristamisel leksikograafias või leksikaal-semantilistes uurimustes, andmekaeves ja masintõlkega seotud ülesannetes. (Wiechmann 2008: 254–257) Eelnevad uurimused on aga tõestanud, et ühe mõõdiku eelistamine teisele on keeruline ja ei saa üheselt öelda, et ühte gruppi kuuluvad statistikud on paremad kui teised (Evert 2008: 32).

Järgnevalt annan ülevaate selles töös rakendatud statistikutest.

2.2.1. Sümmeetrilised mõõdikud

Sümmeetrilised sõnadevahelise seose tugevuse mõõtmise statistikuid võib jagada arvutuslike ja matemaatiliste põhimõtete alusel väiksematesse gruppidesse. Näiteks Evert (2008) on jaganud statistikud lihtsateks (nt t-skoor, vastastikuse informatsiooni väärtus) ja statistilisteks (nt hii-ruut-statistik, log-tõepära funktsioon). Lihtsad statistikud mõõdavad sõnadevahelist seost sõnaühendi sageduse võrdlemisel teoreetilise sagedusega, statistilised mõõdikud põhinevad kahemõõtmelistel sagedustabelitel (vt ptk 1.4.2). Evert (2008: 32) esitab ka teistsuguse liigituse, jagades mõõdikud statistilise mõju suurust (nt vastastikuse informatsiooni väärtus) ja statistilist olulisust (nt t-skoor, hii-ruut-statistik, log-tõepära funktsioon) mõõtvateks statistikuteks.

Selles töös olen sümmeetrilistest statistikutest uurimiseks välja valinud kõige sagedamini sarnaste ülesannete lahendamiseks kasutatavad mõõdikud: lihtne olulisust mõõtev statistik t-skoor (*t-score*) (vt lähemalt Church, Hanks 1990), lihtne efekti suurust mõõtev statistik vastastikuse informatsiooni väärtus (*mutual information*, *MI*) (vt lähemalt Church, Hanks 1990), kaks statistilist olulisust mõõtvat statistikut hii-ruut-statistik (*chi-squared measure*) (vt lähemalt Manning, Schütze 1999) ja log-tõepära funktsioon (*log-likelihood measure*) (vt lähemalt Dunning 1993) ning statistilise mõju suurust mõõtev statistik minimaalne tundlikkus (*minimum sensitivity*, *MS*) (vt lähemalt Pedersen 1998).

2.2.1.1. t-skoor

Lihtsa sõnadevahelise seose tugevuse mõõtmise statistikuna põhineb t-skoor sõnade koosesinemise sageduse võrdlemisel teoreetilise sagedusega. Koosesinemise sagedusest lahutatakse teoreetiline sagedus ning see tulemus jagatakse koosesinemise sageduse ruutjuurega. Valem, kus O on sõnade koosesinemise sagedus valimis ja E on sõnade koosesinemise teoreetiline tõenäosus:

$$\textit{t-skoor:}\ \frac{O-E}{\sqrt{O}}$$

t-skoor on osutunud kasulikuks teatud ülesannete lahendamisel (Evert 2008: 22), näiteks saksa keele prepositsioonifraasi ja verbi ühendite tuvastamisel (Krenn, Evert 2001).

2.2.1.2. Vastastikuse informatsiooni väärtus

Sarnaselt t-skoorile kuulub ka vastastikuse informatsiooni väärtus (MI) lihtsate statistikute rühma ja põhineb sõnaühendi sageduse võrdlemisel teoreetilise sagedusega. Tulemuse saamiseks jagatakse osalausete hulk, kus mõlemad sõnapaari liikmed esinevad, teoreetilise sagedusega, ning sellest võetakse omakorda logaritm alusel kaks. MI arvutatakse valemi põhjal, kus O tähistab sõnade koosesinemise sagedust valimis ja E sõnade koosesinemise teoreetilist tõenäosust:

$$MI = log_2 \frac{O}{E}$$

MI väärtus on kõrgem, kui O on palju suurem kui E, ning tõstab väärtuste loendis kõrgemale harvaesinevaid ühendeid, mille komponendid on samuti madala esinemissagedusega (Evert 2008: 19).

Selleks et parandada MI tulemusi, on Evert (2008: 19) välja pakkunud mitu võimalust. Näiteks võib kogu MI valemi läbi korrutada sõnade koosesinemise sageduse väärtusega. Nii muutub MI väärtuste loend ja kõrgeima MI väärtuse saavad sagedamini esinevad ühendverbid. Teine võimalus on seada sõnapaaride koosesinemise sagedusele piir. Selleks võiks olla näiteks kümme ehk vaatluse alt jäävad välja sõnad, mis esinevad koos samas osalauses vähem kui kümme korda. Selles töös pole ühtegi võimalust MI tulemuste parandamiseks rakendatud.

2.2.1.3. Hii-ruut-statistik

Hii-ruut-statistiku väärtus arvutatakse kahemõõtmelise sagedustabeli väärtuste abil valemi põhjal:

$$hii\text{-}ruut = \sum_{ij} \frac{(Oij - E_{ij})^2}{E_{ij}},$$

$$kus \sum = \frac{(O_{11} - E_{11})^2}{E_{11}} + \frac{(O_{12} - E_{12})^2}{E_{12}} + \frac{(O_{21} - E_{21})^2}{E_{21}} + \frac{(O_{22} - E_{22})^2}{E_{22}}$$

(vt ka ptk 1.4.2).

Hii-ruut-statistik on kahepoolne mõõdik, mis tähendab, et kõrge positiivse väärtuse saavad nii tugeva kui nõrga seosega sõnaühendid. Teisisõnu kõrge statistiku väärtuse saavad nii sõnapaarid, mis kindlasti moodustavad püsiühendi kui ka paarid, mis väldivad koosesinemist. Selleks et kahepoolsest mõõdikust saaks n-ö ühepoolne mõõdik, mis

eristab nii positiivseid kui ka negatiivseid seoseid, korrutatakse tulemus läbi (–1)-ga kui O < E. (Evert 2008: 21)

2.2.1.4. Log-tõepära funktsioon

Log-tõepära funktsioon on kõige laialdasemat kasutust leidnud statistilist olulisust mõõtev statistik eriti arvutilingvistikas (Evert 2008: 31). Sarnaselt hii-ruut-statistikule arvutatakse ka log-tõepära funktsiooni väärtused kahemõõtmelise sagedustabeli väärtuste põhjal:

$$\textit{log-tõepära} = 2\sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}, \textit{ kus summamärk tähistab valimimahtu.}$$

Sarnaselt hii-ruut-statistikule on ka log-tõepära funktsioon kahepoolne mõõdik (Evert 2008: 22) ning vajab ühepoolseks teisendamist (vt eelmine ptk).

2.2.1.5. Minimaalne tundlikkus

Minimaalne tundlikkus (MS) on lihtne ja efektiivne mõõdik bigrammide tuvastamiseks. MS arvutatakse sõna bigrammis esinemise sageduse ja sama sõna üldise esinemissageduse võrdlemisel valemi põhjal:

$$MS = minimum\left(\frac{O11}{O11+O12}, \frac{O11}{O11+O21}\right)$$

Esimene väärtus tähistab esimese sõna tundlikkust teise suhtes ehk juhul kui teine sõna esineb, siis kui suure tõenäosusega esineb esimene sõna. Teine väärtus tähistab vastupidist ehk kui tundlik on teine sõna esimese suhtes ehk kui suure tõenäosusega esineb teine sõna tingimusel, et esineb esimene sõna. Kui kahe väärtuse miinimum on 0, siis kaks vaadeldavat sõna ei esine kunagi koos, kui miinimum on 1, siis on sõnade vahel tugev seos ja iga kord kui üks sõna esineb, esineb ka teine. (Pedersen, Bruce 1996: 12) Daniel Wiechmann (2008: 282) peab MS-i tugevuseks muuhulgas seda, et valimi suuruse muutumine ei mõjuta MS-i tulemust nii suurel määral kui näiteks t-testi, hii-ruut-statistiku või log-tõepära funktsiooni tulemusi. Griesi (2013: 4) järgi on MS-i kasutamine seosetugevuse mõõdikuna mõnevõrra ohtlik, sest kummaski MS-i väärtuses ei avaldu täpselt kumb sõna teise esinemise tingib: $p(word_1|word_2)$ või $p(word_2|word_1)$.

2.2.2. Asümmeetrilised mõõdikud

Kõik eespool kirjeldatud statistikud kajastavad kahe sõna vastastikust seost ja selline lähenemine on korpuslingvistikas kollokatsioonide tuvastamise uurimustes domineerinud

viimased viiskümmend aastat (Gries 2013: 4). Samas on teada, et inimese teadvuses ei ole seos kahe sõna vahel alati sümmeetriline (Michelbacher jt 2007: 1), ning sümmeetrilised mõõdikud ei tuvasta, kas esimene sõna on abiks teise püsiühendi komponendi ennustamisel või vastupidi (Gries 2013: 4). Näiteks ühendi *lahku minema* korral, kui osalauses on *lahku*, siis on seal väga suure tõenäosusega ka *minema* ehk seos *lahku* ja *minema* vahel on tugev ja *minema* esinemine on ennustatav *lahku* järgi, kuid mitte vastupidi: kui osalauses esineb *minema*, siis *lahku* seal sama suure tõenäosusega ei esine ehk seos *minema* ja *lahku* vahel nii tugev ei ole ehk *lahku* ei ole hästi ennustatav *minema* järgi. Erinevalt sümmeetrilistest statistikutest ei ühenda asümmeetrilised statistikud (nt tinglik tõenäosus, ΔP) kahte väga erinevat tõenäosust, vaid arvutavad kaks väärtust: p(sõna$_1$|sõna$_2$) ja p(sõna$_2$|sõna$_1$) (Gries 2013: 4).

Asümmeetrilistest statistikutest testin selles töös tinglikku tõenäosust (*conditional probability*) (vt lähemalt Bell jt 2009; Michelbacher jt 2007) ja ΔP-d (vt lähemalt Ellis 2006; Ellis, Ferreira-Junior 2009).

2.2.2.1. Tinglik tõenäosus

Tinglik tõenäosus on sõnadevahelise seose tugevuse mõõdik, mis on tuletatud MS-ist (vt ptk 2.2.1.5). Kuna tegemist on asümmeetrilise statistikuga, siis tinglik tõenäosus arvutatakse kahe valemi abil:

$$p(word_2|word_1) = \frac{O11}{O11 + O12}$$

$$p(word_1|word_2) = \frac{O11}{O11 + O21}$$

Tinglikku tõenäosust on seose tugevuse mõõdikuna rakendatud vähestes töödes (Gries 2013: 4). Erandiks on Michelbacher jt (2007; 2011) uurimused, mille tulemusena ilmneb, et tinglik tõenäosus on sobiv asümmeetriliste seoste tuvastamiseks, kuid sümmeetriliste seoste tuvastamisel on statistiku tulemuslikkus madal.

2.2.2.2. ΔP

ΔP väärtus arvutatakse kahe tõenäosuse põhjal. Esimese tõenäosuse arvutamisel arvestatakse ennustava sõna sagedust. See tähendab, et võetakse arvesse selle sõna sagedus, mille esinemise abil teise sõna esinemist ennustatakse. Teise tõenäosuse arvutamisel ennustava sõna sagedust ei arvestata. ΔP väärtuse leidmiseks lahutatakse

esimesest tõenäosusest teine. (Ellis 2006: 11) Kahemõõtmelise sagedustabeli abil arvutatakse ΔP väärtused järgmiselt:

$$\Delta P_{2|1} = p(word_2|word_1=esineb) - p(word_2|word_1=puudub) = \frac{O11}{O11+O12} - \frac{O21}{O21+O22}$$

$$\Delta P_{1|2} = p(word_1|word_2=esineb) - p(word_1|word_2=puudub) = \frac{O11}{O11+O21} - \frac{O12}{O12+O22}$$

Kui kahe sündmuse tõenäosused on võrdsed, siis nende sündmuste vahel seos puudub ja ΔP = 0. Kui ΔP = 1, siis teise sõna esinemine suurendab vaadeldava sõna esinemise tõenäosust ja kui ΔP = –1, siis teise sõna esinemine vähendab vaadeldava sõna esinemise tõenäosust ja tegemist on negatiivse seosega. (Ellis 2006: 11)

Gries (2013: 6–13) toob välja ΔP eelised võrreldes sümmeetriliste statistikutega: ΔP on traditsiooniliste mõõdikutega kõrvutades tundlikum, sest vastupidiselt nendele näitab ΔP, missugune sõna kollokatsioonis väljendab tugevamat või nõrgemat seost teiste sõnadega kollokatsioonis; ΔP-d on väga lihtne arvutada; ΔP on intuitiivselt mõistetav; ΔP on leidnud kasutust psühholoogilistest uurimustes ja osutunud mõõdikuks, mis on tugevamini seotud psühholoogilise reaalsusega ning kuna lingvistikat peetakse üha enam kognitiivteaduste osaks, on arvutuslike meetodite kooskõlastamine psühholingvistiliste katsetega eriti oluline.

3. Töö tulemused

Peatükk esitab läbiviidud katsete tulemused. Esimene ja teine alapeatükk kirjeldavad ühendverbide tuvastamisel saadud tulemusi vastavalt erinevaid tekstiliike sisaldavatest ja erineva suurusega korpustest. Kolmandas peatükis esitan sümmeetriliste ja asümmeetriliste mõõdikute võrdluse. Lõpetuseks kõrvutan saadud tulemusi teiste keelte põhjal tehtud sarnaste tööde tulemustega. Kõik arvutused selle töö jaoks tegin statistikaprogrammiga R (R Development CoreTeam 2011).

3.1. Tulemused sõltuvalt tekstiliigist

Peatükk esitab valitud sõnadevahelise seose tugevuse mõõdikute võrdluse aja- ilu- ja teaduskirjandustekstide põhjal. Eesmärk on vaadelda, kas ja kuidas erinevad statistikute tulemused ühendverbide tuvastamisel erinevaid tekstiliike sisaldavatest korpustest.

3.1.1. Lihtsa sagedusloendi täpsus ja saagis olenevalt tekstiliigist

Selleks et vaadelda, kas mõõdikute tulemused sõltuvad tekstiliigist, kasutan Eesti keele koondkorpuse allosa Tasakaalus korpust, mis koosneb 15 miljonist sõnast. Selles korpuses on võrdselt 5 miljonit sõna nii ajalehe-, ilukirjandus- kui ka teadustekste ja on seetõttu tekstiklasside lõikes võrreldavad. Tabel 3 esitab erinevate tekstiklasside andmed: neis sisalduvate osalausete arvu, saadud kandidaatpaaride ja õigete ühendverbide arvud ning sageduse täpsuse ja saagise.

Tabel 3. **Sageduse täpsus ja saagis õigete ühendverbide tuvastamisel Tasakaalus korpusest.**

korpus	osalaused	kandidaatpaarid	õiged ühendverbid	täpsus	saagis
ajakirjandus	707 979	13141	1351	10,3%	77,8%
ilukirjandus	912 101	18459	1519	8,2%	87,4%
teaduskirjandus	677 756	8218	972	11,8%	56,0%

5 miljonit sõna ilukirjandustekste sisaldab 912 101 osalauset, mida on rohkem, kui teisi tekstiliike sisaldavates korpustes. Ilukirjandustekstide põhjal leitakse ka kõige rohkem EKSS-is esitatud ühendverbidest (87,4%). Samas on ilukirjandustekstidest ühendverbide tuvastamisel sageduse täpsus kõige madalam (8,2%), sest genereeriti kõige rohkem kandidaatpaare. Teadustekste võib aga pidada ühendverbide automaattuvastamise kontekstis kõige problemaatilisemaks tekstiklassiks, sest sagedus tuvastab kokku vaid 56,0% EKSS-i nimistus olevatest ühendverbidest. Samas on teaduskirjandustekstidest leitud ühenditest 11,8% EKSS-is ja teadustekstidest genereeritud kandidaatpaaride seas on n-ö müra kõige vähem.

3.1.2. Sümmeetriliste mõõdikute ja lihtsa sagedusloendi tulemused Tasakaalus korpuse põhjal

See peatükk annab ülevaate sümmeetriliste statistikute ja lihtsa sagedusloendi tulemustest aja-, ilu- ja teaduskirjandustekstide lõikes. Eesmärk on võrrelda valitud meetodite tulemusi erinevate tekstiklasside peal.

3.1.2.1. t-skoori tulemused Tasakaalus korpuse põhjal

Tabel 4[7] esitab kahanevalt 50 kõrgeima t-skoori väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.

Tabel 4. **50 kõrgeima t-skoori väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; t-skoor – t-skoori väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**t-skoor**	**adverb_ verb**	**sag**	**yv**	**t-skoor**	**adverb_ verb**	**sag**	**yv**	**t-skoor**
vastu_ võtma	793	T	27,144	otsa_ vaatama	1034	F	31,480	ette_ nägema	843	T	28,638
ette_ nägema	780	T	26,879	tagasi_ tulema	1038	T	28,387	kaasa_ tooma	760	T	27,091
välja_ tulema	812	T	21,042	ette_ kujutama	804	T	28,153	välja_ tooma	817	T	27,006
kaasa_ tooma	402	T	19,416	välja_ tulema	1184	T	26,197	läbi_ viima	732	T	26,829
välja_ andma	557	T	18,521	ära_ tundma	838	T	25,312	välja_ töötama	682	T	25,366

[7] Kõik tabelid, mis ei ole töös esitatud täielikul kujul, on leitavad aadressil http://kodu.ut.ee/~eleriaed/magistrit88_failid/.

alla_ kirjutama	352	T	18,274	vastu_ võtma	631	T	23,279	esile_ tooma	573	T	23,420
kinni_ pidama	412	T	17,894	kinni_ hoidma	558	T	23,003	kaasa_ aitama	388	T	19,482
välja_ kuulutama	338	T	17,813	kaasa_ võtma	562	F	22,254	vastu_ võtma	379	T	19,138
ette_ kujutama	320	T	17,691	püsti_ tõusma	437	F	20,704	välja_ selgitama	381	T	18,887
kokku_ leppima	322	T	17,689	lahti_ tegema	568	T	20,467	välja_ andma	501	T	18,443
ette_ võtma	393	T	17,540	kinni_ võtma	507	T	19,874	välja_ kujunema	344	T	17,737
tagasi_ tulema	420	T	16,887	välja_ nägema	670	T	19,290	esile_ kutsuma	316	T	17,676
läbi_ viima	302	T	16,662	sisse_ astuma	385	T	18,941	esile_ tõstma	310	T	17,485
ette_ valmistama	286	T	16,635	ette_ võtma	457	T	18,506	välja_ pakkuma	312	T	16,597
maha_ müüma	255	T	15,549	maha_ jätma	376	T	18,331	üle_ minema	283	T	16,379
kaasa_ aitama	255	T	15,541	ringi_ vaatama	376	T	18,120	välja_ arvama	274	T	14,839
alla_ jääma	282	T	15,313	kinni_ panema	396	T	17,922	kokku_ langema	225	T	14,762
üle_ andma	321	T	15,050	üles_ leidma	331	T	17,059	ära_ tundma	228	T	14,643
ilma_ jääma	243	T	14,966	ära_ võtma	562	T	16,937	ette_ valmistama	189	T	13,586
maha_ võtma	272	T	14,683	tagasi_ minema	492	T	16,705	kokku_ puutuma	183	T	13,454
vastu_ pidama	329	T	14,676	ära_ minema	698	T	16,646	ära_ tooma	210	T	13,399
välja_ tooma	307	T	14,580	peale_ hakkama	334	T	16,633	esile_ tulema	211	F	12,857
kaasa_ võtma	267	F	14,575	maha_ laskma	321	T	16,526	ära_ kasutama	228	T	12,773
ära_ kasutama	265	T	14,503	lahti_ laskma	300	T	16,125	kokku_ leppima	159	T	12,537
välja_ nägema	343	T	14,431	vastu_ pidama	390	T	16,068	ette_ heitma	155	T	12,381
valmis_ olema	614	F	14,189	üle_ jääma	349	T	15,794	kokku_ võtma	187	T	12,253
kinni_ maksma	219	T	14,090	läbi_ käima	308	T	15,736	edasi_ andma	183	T	12,222
välja_ pakkuma	263	T	13,912	kokku_ saama	449	T	15,652	välja_ jääma	235	T	12,193
ära_ võtma	319	T	13,820	ringi_ käima	262	T	15,231	ära_ hoidma	152	T	12,117
välja_ töötama	235	T	13,323	edasi_ minema	387	T	15,093	välja_ jätma	164	T	11,810
edasi_ lükkama	180	T	13,209	kinni_ haarama	237	F	14,978	üles_ ehitama	133	T	11,482
välja_ valima	231	T	13,090	üles_ otsima	234	T	14,512	alla_ jääma	140	T	11,349

edasi_ minema	238	T	13,076	üles_ tõstma	228	T	14,445	kõrvale_ jätma	130	T	11,333
kokku_ hoidma	191	T	12,814	välja_ astuma	294	T	14,410	üle_ võtma	151	T	10,843
ära_ tegema	387	T	12,675	vastu_ vaidlema	209	T	14,328	välja_ tulema	261	T	10,697
maha_ jääma	222	T	12,632	kaasa_ tooma	220	T	14,222	ette_ tulema	166	T	10,567
vahele_ jääma	160	T	12,308	lahti_ saama	400	T	14,097	üle_ kandma	112	T	10,172
välja_ vahetama	171	T	12,297	ära_ sööma	247	T	13,799	kaasa_ arvama	119	T	10,151
üle_ jääma	251	T	12,261	kokku_ leppima	183	T	13,355	sisse_ tooma	107	T	9,958
kokku_ puutuma	153	T	12,179	järele_ mõtlema	213	T	13,291	ette_ kujutama	103	T	9,804
läbi_ käima	185	T	12,146	ette_ nägema	284	T	13,255	esile_ kerkima	97	F	9,776
tagasi_ lükkama	150	T	11,948	välja_ mõtlema	388	T	13,126	välja_ valima	130	T	9,724
tagasi_ astuma	154	T	11,859	valmis_ olema	548	F	13,121	edasi_ arendama	96	F	9,686
esile_ tõstma	136	T	11,582	mööda_ minema	221	T	13,100	välja_ arendama	102	T	9,556
üles_ kutsuma	144	T	11,568	maha_ võtma	283	T	13,051	välja_ kasvama	121	T	9,534
kinni_ hoidma	143	T	11,512	ära_ viima	237	T	12,963	maha_ jääma	97	T	9,477
välja_ mõtlema	201	T	11,389	kokku_ puutuma	174	T	12,946	tagasi_ tulema	109	T	9,329
ette_ heitma	135	T	11,385	üle_ minema	327	T	12,936	kinni_ pidama	95	T	9,140
üles_ astuma	138	T	11,373	kinni_ pidama	327	T	12,910	ära_ määrama	99	T	9,033
maha_ laskma	142	T	11,298	maha_ müüma	174	T	12,895	esile_ tõusma	82	F	8,800

Tabelist 4 on näha, et t-skoor tuvastab suure hulga õigeid ühendverbe. Ajakirjandustekstide põhjal leitud 50-st kõrgeima t-skoori väärtuse saanud sõnapaarist vaid kaks ei kuulu EKSS-i loendisse (*kaasa võtma*, *valmis olema*). Ilu- ja teaduskirjandustekstide 50 kõrgeima t-skoori väärtusega ühendi hulgas on vastavalt viis ja neli EKSS-i nimistusse mitte kuuluvat ühendit. Ajakirjandustekstides on need *otsa vaatama*, *kaasa võtma*, *püsti tõusma*, *kinni haarama*, *valmis olema*, teaduskirjanduses aga *esile tulema*, *esile kerkima*, *edasi arendama* ja *esile tõusma*.

t-skoor töötab kõikide tekstiklasside peal väga hästi – enamik tuvastatud ühenditest on kõrge sagedusega õiged ühendverbid. Teised tuvastatud ühendid küll EKSS-i

loendisse ei kuulu, kuid on siiski korpuses sagedalt esinevad ja ilmselgelt keeles ühendverbidena esinevad.

3.1.2.2. MI tulemused Tasakaalus korpuse põhjal

Tabel 5 esitab kahanevalt 50 kõrgeima MI väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.

Tabel 5. **50 kõrgeima MI väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; MI – MI väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**MI**	**adverb_ verb**	**sag**	**yv**	**MI**	**adverb_ verb**	**sag**	**yv**	**MI**
sekka_ hällitama	1	F	14,04	ülal_ ekshibeerima	1	F	12,87	valla_ pääsima	1	F	14,91
tagant_ suskima	1	F	13,32	tagant_ punima	1	F	12,62	järel_ vantsima	1	F	14,56
ühte_ lõpma	1	F	12,82	tagant_ utjama	1	F	12,62	üleval_ semantiseeruma	1	F	14,24
tagant_ piitsutama	3	T	11,59	valla_ loppuma	1	F	12,57	vahele_ tükkima	1	F	13,88
külge_ kleepuma	3	T	11,54	lahku_ translitereerima	1	F	12,50	taga_ lähnuma	1	F	13,70
külge_ pookima	3	T	11,37	peal_ ruiguma	1	F	12,32	külge_ liimima	2	F	13,66
püsti_ krapsama	1	F	11,28	kallale_ hassetama	1	F	12,12	ühte_ kollabseeruma	1	F	13,59
laiali_ hoorama	1	F	11,27	alt_ kolmandama	1	F	11,91	külge_ kaema	2	F	13,46
vahele_ pläristama	1	F	11,27	alt_ muduma	1	F	11,91	püsti_ kohisema	1	F	13,35
peal_ kondama	1	F	11,05	tagant_ pandsima	1	F	11,62	vahele_ käristama	1	F	12,88
peal_ parvetama	1	F	11,05	valla_ nukerdama	1	F	11,57	pealt_ kaema	3	F	12,79
tagant_ suskama	1	F	11,00	üleval_ poositama	1	F	11,38	taga_ kiusama	2	T	12,70
kaasas_ tilpnema	1	F	10,94	peal_ kriipsima	1	F	11,32	taga_ tõttama	1	F	12,70
kõrvalt_ kiristama	1	F	10,86	peal_ tsinkima	1	F	11,32	pealt_ hammustama	1	F	12,62
mööda_ sudima	1	F	10,68	pealt_ ilgema	1	F	11,26	kallale_ tungima	2	F	12,37
peal_ lupjama	1	F	10,64	pealt_ satuma	1	F	11,26	mööda_ visioneerima	1	F	12,37
pealt_ äsama	1	F	10,63	tagant_ turmama	1	F	11,03	püsti_ loivama	1	F	12,35
sekka_ köhima	1	F	10,58	valla_ oppama	1	F	10,99	tagant_ kütma	2	F	12,32

tagant_ togima	1	F	10,52	kaasas_ taastunnustama	1	F	10,85	järele_ ahvima	1	F	12,04
järele_ reformeerima	1	F	10,47	külge_ pookima	5	T	10,72	järele_ võsastuma	1	F	12,04
täis_ maruma	1	F	10,28	tasa_ muisklema	1	F	10,65	järele_ önnistuma	1	F	12,04
täis_ roojama	1	F	10,28	tagant_ mehitama	1	F	10,62	tagant_ tõukama	4	T	12,04
eemale_ peletama	34	F	10,27	järel_ krousima	1	F	10,57	ringi_ meikima	1	F	12,03
laiali_ pillutama	4	F	10,27	kõrvalt_ sandistama	1	F	10,45	ringi_ pöruma	1	F	12,03
laiali_ pihustuma	1	F	10,27	laiali_ sõõrduma	5	F	10,45	ringi_ sobrama	1	F	12,03
ilma_ etiketistama	1	F	10,10	laiali_ kluksuma	1	F	10,45	ringi_ vurama	1	F	12,03
ilma_ kekutama	1	F	10,10	laiali_ siia-sinama	1	F	10,45	peal_ kosima	1	F	12,00
taga_ protsessima	1	F	10,10	alt_ oopima	1	F	10,33	taha_ paisuma	1	F	11,88
peal_ koperdama	1	F	10,05	alt_ tarnima	1	F	10,33	tagant_ kannustama	1	F	11,80
ringi_ floresiensima	1	F	10,04	peal_ köietama	1	F	10,32	laiali_ pillutama	3	F	11,77
ringi_ kooberdama	1	F	10,04	eemale_ lootsima	1	F	10,06	püsti_ kükitama	1	F	11,76
ringi_ melutsema	1	F	10,04	valmis_ schoolima	1	F	10,02	püsti_ pannema	1	F	11,76
ringi_ roobeldama	1	F	10,04	taha_ kõkutama	1	F	9,99	lahti_ rulluma	3	T	11,73
taga_ kiusama	10	T	9,96	valla_ prevaleerima	1	F	9,99	lahti_ muukima	2	F	11,73
järel_ lohisema	1	F	9,88	tagant_ surkima	2	F	9,92	lahti_ kasukkama	1	F	11,73
ümber_ grupeeruma	1	F	9,83	taha_ loovima	1	F	9,90	lahti_ korkima	1	T	11,73
üleval_ julgestama	1	F	9,83	ilma_ jõutama	1	F	9,87	taga_ haukuma	1	F	11,70
koos_ armeerima	1	F	9,79	taga_ haspeldama	1	F	9,86	eemale_ peletama	7	F	11,53
koos_ kudrutama	1	F	9,79	tagant_ utsitama	1	T	9,81	lahti_ harutama	5	T	11,47
koos_ kypsetama	1	F	9,79	sekka_ lükkima	1	F	9,75	tasa_ akumuleerima	1	F	11,46
tagant_ utsitama	1	T	9,74	peal_ veretama	1	F	9,74	ringi_ uitama	2	F	11,45
kõrvalt_ seirama	1	F	9,73	mööda_ ehituma	1	F	9,67	taga_ hõikama	1	F	11,38
tagant_ torkima	3	T	9,70	mööda_ neli-viima	1	F	9,67	taga_ kargama	1	F	11,38
mööda_ loovima	2	F	9,68	mööda_ vuhkima	1	F	9,67	külge_ valetama	1	F	11,38
mööda_ logistama	1	F	9,68	kallale_ ässitama	8	F	9,66	mööda_ laveerima	1	F	11,37

otsa_ elektritarvittima	1	F	9,66	ümber_ plartseldama	2	F	9,63	mööda_ libistama	1	F	11,37
peal_ mürisema	1	F	9,64	ümber_ defineeruma	1	F	9,63	mööda_ pügama	1	F	11,37
peal_ treima	1	F	9,64	ümber_ parantsatama	1	F	9,63	mööda_ seilama	1	F	11,37
pealt_ putitama	1	F	9,63	ümber_ plumpsuma	1	F	9,63	tasa_ viluma	1	F	11,35
valmis_ jäigendama	1	F	9,59	ümber_ profileerima	1	F	9,63	kinni_ tukastama	2	F	11,32

Erinevalt t-skoori tulemustest on MI kõrgemad väärtused saanud ühendid, mis EKSS-i loendisse ei kuulu. Nagu eespool mainitud, tõstab MI harvaesinevaid ühendeid, mille komponentide esinemissagedus on samuti madal. Seda kinnitab asjaolu, et 50-st kõrgeima MI väärtuse saanud ühendist ajakirjandustekstides ja teaduskirjanduses on vastavalt kuus ja viis õiged ühendverbid (nt *külge kleepuma*, *tagant utsitama*, *lahti harutama*), mille sagedus on madal. Kõige vähem õigeid ühendverbe 50 kõrgeima MI väärtusega ühendite seas on ilukirjandustekstides, kus neid on kaks (*külge pookima*, *tagant utsitama*).

MI tulemused on kõikide tekstiliikide põhjal halvad – loendite eesotsas on ühendid, mis esinevad alamkorpuses kuni kolm korda, sageli vaid korra. MI saab ühendverbide tuvastamisega paremini hakkama aja- ja teaduskirjandusest, halvemini ilukirjandustekstidest.

3.1.2.3. Hii-ruut-statistiku tulemused Tasakaalus korpuse põhjal

Tabel 6 esitab kahanevalt 50 kõrgeima hii-ruut-statistiku väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides (väärtused on läbi korrutatud (–1)-ga, juhul kui O < E).

Tabel 6. **50 kõrgeima hii-ruut-statistiku väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.** Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; hii-ruut – hii-ruut-statistiku väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**hii-ruut**	**adverb_ verb**	**sag**	**yv**	**hii-ruut**	**adverb_ verb**	**sag**	**yv**	**hii-ruut**
eemale_ peletama	34	F	42051	ette_ kujutama	804	T	112007	läbi_ viima	732	T	86586
ette_ kujutama	320	T	28566	otsa_ vaatama	1034	F	47823	ette_ nägema	843	T	60613
kokku_ leppima	322	T	22103	püsti_ tõusma	437	F	45006	esile_ kutsuma	316	T	55259

alt_ vedama	55	T	21244	eemale_ peletama	76	F	40652	esile_ tõstma	310	T	44324
vastu_ võtma	793	T	20763	vastu_ vaidlema	209	T	23170	kaasa_ tooma	760	T	42885
esile_ tõstma	136	T	19543	kinni_ hoidma	558	T	20349	lahti_ mõtestama	37	T	36948
ette_ nägema	780	T	19486	alla_ neelama	114	T	14959	kaasa_ aitama	388	T	34858
ette_ valmistama	286	T	17040	kokku_ leppima	183	T	14013	kokku_ puutuma	183	T	33190
sekka_ hällitama	1	F	16856	pärale_ jõudma	98	T	13716	valla_ pääsima	1	F	30806
alla_ kirjutama	352	T	12948	vahele_ segama	66	T	10825	üles_ ehitama	133	T	30097
kaasa_ tooma	402	T	12031	sisse_ astuma	385	T	10433	ette_ heitma	155	T	27778
kokku_ põrkama	103	T	11909	püsti_ kargama	86	F	10397	kokku_ leppima	159	T	27325
kallale_ tungima	36	F	11683	taga_ ajama	143	T	10030	külge_ liimima	2	F	25816
edasi_ lükkama	180	T	11312	kokku_ puutuma	174	T	9083	esile_ tooma	573	T	25607
välja_ kuulutama	338	T	10338	laiali_ valguma	38	F	8653	järel_ vantsima	1	F	24205
tagant_ suskima	1	F	10260	tagant_ õhutama	11	F	8532	välja_ töötama	682	T	22697
taga_ kiusama	10	T	9962	külge_ pookima	5	T	8438	külge_ kaema	2	F	22589
ümber_ lükkama	97	T	9913	kinni_ haarama	237	F	8339	vastu_ võtma	379	T	21811
kokku_ puutuma	153	T	9705	kaasa_ võtma	562	F	8221	kõrvale_ jätma	130	T	21443
maha_ müüma	255	T	9255	kallale_ tungima	30	F	7971	pealt_ kaema	3	F	21175
tagant_ piitsutama	3	T	9230	kõrvale_ põikama	27	F	7786	eemale_ peletama	7	F	20681
kaasa_ aitama	255	T	9079	ühte_ sulama	18	T	7540	üleval_ semantiseeruma	1	F	19363
külge_ kleepuma	3	T	8944	vastu_ võtma	631	T	7533	tagant_ tõukama	4	T	16780
edasi_ lükkuma	44	F	8463	ülal_ ekshibeerima	1	F	7475	tagasi_ lükkama	68	T	16552
esile_ kutsuma	105	T	8406	maha_ müüma	174	T	7447	ette_ valmistama	189	T	15764
külge_ pookima	3	T	7950	ette_ valmistama	138	T	7434	vahele_ tükkima	1	F	15060
ühte_ lõpma	1	F	7223	laiali_ sõõrduma	5	F	7000	kaasas_ käima	51	T	15025
läbi_ viima	302	T	6797	tagasi_ tulema	1038	T	6983	ühte_ sulama	9	T	14663
ette_ heitma	135	T	6464	sisse_ hingama	131	T	6892	alla_ kriipsutama	15	T	14509
kinni_ nabima	19	T	6339	esile_ kutsuma	75	T	6583	lahti_ harutama	5	T	14185
peale_ suruma	62	T	6216	lahti_ harutama	46	T	6526	kokku_ langema	225	T	13798

kõrvale_ hiilima	16	T	6210	kokku_ varisema	75	T	6489	välja_ tooma	817	T	13435
tagasi_ lükkama	150	T	5868	kallale_ ässitama	8	F	6457	taga_ lähnuma	1	F	13288
taga_ ajama	50	T	5790	külge_ kleepima	12	T	6393	taga_ kiusama	2	T	13286
vahele_ jääma	160	T	5706	laiali_ pillutama	9	F	6291	esile_ kerkima	97	F	12981
ilma_ jääma	243	T	5704	tagant_ punima	1	F	6289	ühte_ kollabseeruma	1	F	12322
otsa_ sõitma	105	T	5365	tagant_ utjama	1	F	6289	välja_ selgitama	381	T	11129
laiali_ pillutama	4	F	4928	maha_ jätma	376	T	6210	kallale_ tungima	2	F	10588
lahti_ rebima	33	F	4796	kokku_ korjama	116	T	6159	edasi_ lükkama	67	T	10508
välja_ lülitama	107	T	4613	eemale_ tõukama	41	T	6081	laiali_ pillutama	3	F	10440
esile_ kerkima	38	F	4417	valla_ loppuma	1	F	6080	püsti_ kohisema	1	F	10426
pärale_ jõudma	32	T	4345	üles_ pooma	59	T	6019	kõrvale_ kalduma	39	T	10345
lahti_ harutama	10	T	4325	lahku_ translitereerima	1	F	5809	üle_ minema	283	T	10235
lahti_ mõtestama	18	T	4187	valla_ päästma	25	T	5722	lahti_ rulluma	3	T	10214
kinni_ maksma	219	T	4179	ringi_ kolama	32	F	5695	tagant_ kütma	2	F	10188
vastu_ vaidlema	55	T	4098	püsti_ ajama	175	F	5624	tagasi_ pöörduma	75	F	8856
üles_ astuma	138	T	4082	sisse_ lülitama	74	T	5322	maha_ suruma	29	T	8785
alla_ neelama	33	T	4072	ära_ tundma	838	T	5213	ära_ hoidma	152	T	8580
kõrvale_ kalduma	20	T	3948	püsti_ hüppama	76	F	5203	edasi_ arendama	96	F	8214
üles_ ehitama	136	T	3932	peal_ ruiguma	1	F	5123	vahele_ käristama	1	F	7529

Tabelist 6 selgub, et hii-ruut-statistiku tulemused ajakirjandustekstidest õigete ühendverbide tuvastamisel on paremad kui teistest tekstiklassidest – ajakirjandustekstide 50-st kõrgeima hii-ruut-statistiku väärtusega ühendist on 41 õiged ühendverbid, ilu- ja teaduskirjandustekstide ühendite seas on see arv vastavalt 28 ja 32. Kõikides tekstiklassides on kõrge hii-ruut-statistiku väärtuse saanud ühendite hulgas nii kõrge sagedusega ühendeid (nt *vastu võtma*, *otsa vaatama*, *välja tooma*) kui ka selliseid, mis on korpuses vaid korra esinenud (nt *tagant suskima*, *tagant utjama*, *järel vantsima*).

Seega hii-ruut-statistiku tuvastatud ühendite seas on igas tekstiklassis nii kõrge sagedusega õigeid ühendverbe kui ka korra korpuses esinenud ühendeid. 50 kõrgeima hii-

ruut-statistiku väärtuse saanud ühendi põhjal saab väita, et kõige paremini tuvastab mõõdik ühendverbe ajakirjandustekstidest, halvemini teadus- ja ilukirjandustekstidest. Võrreldes MI-ga on hii-ruut-statistiku tulemused iga tekstiliigi korral märkimisväärselt paremad.

3.1.2.4. Log-tõepära funktsiooni tulemused Tasakaalus korpuse põhjal

Tabel 7 esitab kahanevalt 50 kõrgeima log-tõepära funktsiooni väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides (väärtused on läbi korrutatud (–1)-ga, juhul kui O < E).

Tabel 7. **50 kõrgeima log-tõepära funktsiooni väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; log-t – log-tõepära funktsiooni väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**log-t**	**adverb_ verb**	**sag**	**yv**	**log-t**	**adverb_ verb**	**sag**	**yv**	**log-t**
vastu_ võtma	793	T	4120,502	ette_ kujutama	804	T	7207,088	ette_ nägema	843	T	6180,867
ette_ nägema	780	T	3895,970	otsa_ vaatama	1034	F	6843,053	läbi_ viima	732	T	6165,115
ette_ kujutama	320	T	2445,957	püsti_ tõusma	437	F	3431,758	kaasa_ tooma	760	T	5178,161
kokku_ leppima	322	T	2287,497	kinni_ hoidma	558	T	3149,768	välja_ töötama	682	T	3816,645
kaasa_ tooma	402	T	2111,417	tagasi_ tulema	1038	T	2808,349	esile_ tooma	573	T	3558,923
alla_ kirjutama	352	T	1992,026	vastu_ võtma	631	T	2272,574	välja_ tooma	817	T	3419,837
ette_ valmistama	286	T	1905,995	kaasa_ võtma	562	F	2234,634	kaasa_ aitama	388	T	2925,330
välja_ kuulutama	338	T	1852,942	ära_ tundma	838	T	2126,668	esile_ kutsuma	316	T	2861,295
maha_ müüma	255	T	1418,347	sisse_ astuma	385	T	1928,372	esile_ tõstma	310	T	2650,799
kaasa_ aitama	255	T	1408,661	välja_ tulema	1184	T	1701,135	vastu_ võtma	379	T	2588,774
läbi_ viima	302	T	1406,666	vastu_ vaidlema	209	T	1667,017	välja_ selgitama	381	T	2005,966
ilma_ jääma	243	T	1202,846	maha_ jätma	376	T	1535,758	kokku_ puutuma	183	T	1709,995
edasi_ lükkama	180	T	1199,836	ringi_ vaatama	376	T	1426,323	välja_ kujunema	344	T	1582,068
esile_ tõstma	136	T	1139,503	kinni_ võtma	507	T	1344,850	üle_ minema	283	T	1577,143
kinni_ pidama	412	T	1126,354	lahti_ tegema	568	T	1330,910	kokku_ langema	225	T	1495,101

ette_ võtma	393	T	1053,572	kokku_ leppima	183	T	1305,325	kokku_ leppima	159	T	1457,679
kokku_ puutuma	153	T	1051,098	kinni_ haarama	237	F	1302,521	ette_ heitma	155	T	1416,598
välja_ tulema	812	T	1043,545	üles_ leidma	331	T	1262,422	ette_ valmistama	189	T	1374,854
vahele_ jääma	160	T	955,312	kinni_ panema	396	T	1162,584	üles_ ehitama	133	T	1240,543
kinni_ maksma	219	T	951,107	ette_ võtma	457	T	1094,630	välja_ pakkuma	312	T	1222,224
alla_ jääma	282	T	905,008	kokku_ puutuma	174	T	1093,016	ära_ tundma	228	T	1199,836
kokku_ põrkama	103	T	881,796	maha_ laskma	321	T	1086,791	kõrvale_ jätma	130	T	1136,534
välja_ andma	557	T	879,937	lahti_ laskma	300	T	1083,215	ära_ hoidma	152	T	975,475
tagasi_ lükkama	150	T	854,669	peale_ hakkama	334	T	1055,923	välja_ andma	501	T	962,533
ette_ heitma	135	T	828,607	ringi_ käima	262	T	1031,272	esile_ kerkima	97	F	800,513
tagasi_ tulema	420	T	816,012	üles_ tõstma	228	T	1029,071	välja_ arvama	274	T	781,280
maha_ võtma	272	T	751,734	maha_ müüma	174	T	1025,967	ära_ tooma	210	T	723,880
esile_ kutsuma	105	T	747,129	kaasa_ tooma	220	T	1014,379	edasi_ arendama	96	F	691,281
kaasa_ võtma	267	F	745,679	taga_ ajama	143	T	990,963	alla_ jääma	140	T	657,942
ümber_ lükkama	97	T	732,505	üles_ otsima	234	T	976,673	tagasi_ lükkama	68	T	639,812
ära_ kasutama	265	T	727,869	pärale_ jõudma	98	T	945,616	tagasi_ pöörduma	75	F	587,178
üles_ astuma	138	T	705,107	alla_ neelama	114	T	937,956	edasi_ lükkama	67	T	569,717
üles_ kutsuma	144	T	700,650	püsti_ ajama	175	F	925,674	esile_ tulema	211	F	565,815
üles_ ehitama	136	T	688,798	ette_ valmistama	138	T	879,899	välja_ jätma	164	T	555,272
tagasi_ astuma	154	T	686,673	läbi_ käima	308	T	876,134	edasi_ andma	183	T	552,391
kinni_ hoidma	143	T	685,543	välja_ nägema	670	T	875,203	kokku_ võtma	187	T	528,990
välja_ vahetama	171	T	680,109	eemale_ peletama	76	F	857,383	üle_ kandma	112	T	526,611
välja_ lülitama	107	T	676,820	sisse_ hingama	131	T	822,006	sisse_ tooma	107	T	521,850
kokku_ hoidma	191	T	668,442	välja_ lülitama	136	T	809,311	ette_ kujutama	103	T	511,141
üle_ andma	321	T	665,033	üles_ ärkama	153	T	796,946	kaasas_ käima	51	T	508,486
otsa_ sõitma	105	T	656,927	üle_ jääma	349	T	743,171	ära_ kasutama	228	T	489,298
välja_ tooma	307	T	608,657	kokku_ korjama	116	T	731,085	maha_ jääma	97	T	471,815

tagasi_ pöörduma	114	F	608,503	vastu_ pidama	390	T	714,446	välja_ kuulutama	80	T	470,158
kaasa_ lööma	123	T	603,533	koos_ elama	158	F	681,184	lahti_ mõtestama	37	T	456,693
välja_ pakkuma	263	T	598,569	püsti_ kargama	86	F	677,878	esile_ tõusma	82	F	436,722
vastu_ pidama	329	T	596,300	tagasi_ minema	492	T	669,860	kaasa_ arvama	119	T	424,418
maha_ laskma	142	T	587,091	järele_ mõtlema	213	T	667,682	välja_ arendama	102	T	420,494
alt_ vedama	55	T	578,219	edasi_ liikuma	156	F	652,037	üle_ võtma	151	T	391,712
valmis_ olema	614	F	576,134	ära_ võtma	562	T	633,316	kinni_ pidama	95	T	384,717
välja_ töötama	235	T	569,701	maha_ lööma	191	T	630,301	välja_ jääma	235	T	384,471

50-st ajakirjandustekstidest tuvastatud kõrgeima log-tõepära funktsiooni väärtusega ühendist ei kuulu kolm EKSS-i loendisse (*kaasa võtma*, *tagasi pöörduma*, *valmis olema*). See tulemus on võrreldav t-skoori tulemusega. Samas ilukirjanduses on valede ühendverbide (nt *kinni haarama*, *püsti kargama*) arv üheksa, teaduskirjanduses viis (nt *esile kerkima*, *edasi arendama*), mida on rohkem kui t-skoori tuvastatud ühendites. Kõige paremini saab log-tõepära funktsioon hakkama ajakirjandustekstidest ühendverbide tuvastamisega, kõige halvemini aga ilukirjandustekstidest.

Log-tõepära funktsiooni võib ühendverbide tuvastamisel igast tekstiklassist pidada pisut halvemaks meetodiks t-skoorist, kuid selgelt paremaks hii-ruut-statistikust ja MI-st.

3.1.2.5. MS-i tulemused Tasakaalus korpuse põhjal

Tabel 8 esitab kahanevalt 50 kõrgeima MS-i väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.

Tabel 8. **50 kõrgeima MS-i väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F =ei ole õige ühendverb; MS – MS-i väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**MS**	**adverb_ verb**	**sag**	**yv**	**MS**	**adverb_ verb**	**sag**	**yv**	**MS**
eemale_ peletama	34	F	0,142	ette_ kujutama	804	T	0,191	läbi_ viima	732	T	0,296
ette_ nägema	780	T	0,128	püsti_ tõusma	437	F	0,146	kaasa_ aitama	388	T	0,226

kaasa_ aitama	255	T	0,115	kinni_ hoidma	558	T	0,137	ette_ nägema	843	T	0,215
ümber_ lükkama	97	T	0,115	sisse_ astuma	385	T	0,107	esile_ kutsuma	316	T	0,195
maha_ müüma	255	T	0,113	otsa_ vaatama	1034	F	0,089	esile_ tõstma	310	T	0,192
esile_ tõstma	136	T	0,100	eemale_ peletama	76	F	0,089	lahti_ mõtestama	37	T	0,186
kallale_ tungima	36	F	0,100	maha_ jätma	376	T	0,086	üles_ ehitama	133	T	0,179
kaasa_ tooma	402	T	0,097	vahele_ segama	66	T	0,081	kaasa_ tooma	760	T	0,146
ette_ kujutama	320	T	0,094	ära_ tundma	838	T	0,076	välja_ tooma	817	T	0,139
alla_ kirjutama	352	T	0,090	tagant_ õhutama	11	F	0,076	külge_ liimima	2	F	0,133
läbi_ viima	302	T	0,090	üles_ leidma	331	T	0,069	külge_ kaema	2	F	0,133
vastu_ võtma	793	T	0,084	püsti_ kargama	86	F	0,068	ühte_ sulama	9	T	0,132
ette_ valmistama	286	T	0,084	ühte_ sulama	18	T	0,066	üle_ minema	283	T	0,121
esile_ kerkima	38	F	0,078	vastu_ vaidlema	209	T	0,066	tagant_ tõukama	4	T	0,118
edasi_ lükkama	180	T	0,077	kaasa_ tooma	220	T	0,063	välja_ töötama	682	T	0,116
kokku_ leppima	322	T	0,074	tagant_ paisuma	9	F	0,062	esile_ tooma	573	T	0,110
üles_ kutsuma	144	T	0,073	maha_ laskma	321	T	0,061	kõrvale_ kalduma	39	T	0,108
üles_ astuma	138	T	0,070	kallale_ tungima	30	F	0,060	tagant_ kütma	2	F	0,105
alt_ vedama	55	T	0,069	üles_ otsima	234	T	0,060	ära_ tundma	228	T	0,101
üles_ ehitama	136	T	0,069	püsti_ hüppama	76	F	0,060	tagasi_ pöörduma	75	F	0,097
kinni_ hoidma	143	T	0,063	laiali_ valguma	38	F	0,058	kokku_ langema	225	T	0,095
maha_ laskma	142	T	0,061	üles_ tõstma	228	T	0,058	ette_ valmistama	189	T	0,095
peale_ suruma	62	T	0,060	kinni_ haarama	237	F	0,058	kõrvale_ jätma	130	T	0,088
välja_ andma	557	T	0,058	lahti_ laskma	300	T	0,057	tagasi_ lükkama	68	T	0,088
kaasa_ lööma	123	T	0,057	ümber_ pöörama	82	T	0,056	ligi_ pääsema	24	T	0,084
esile_ kutsuma	105	T	0,056	ümber_ pöörduma	64	T	0,056	pealt_ kaema	3	F	0,083
külge_ kleepima	5	T	0,056	tagant_ torkima	8	T	0,055	edasi_ arendama	96	F	0,081
taga_ nutma	12	T	0,056	välja_ tulema	1184	T	0,054	ette_ heitma	155	T	0,078
eemale_ tõrjuma	13	F	0,052	eemale_ tõmbuma	46	F	0,054	kokku_ puutuma	183	T	0,077

alla_ kukkuma	84	F	0,051	külge_ kleepima	12	T	0,053	ümber_ lükkama	35	T	0,076
tagasi_ astuma	154	T	0,050	kõrvale_ lükkama	60	T	0,052	ära_ hoidma	152	T	0,074
kinni_ maksma	219	T	0,049	välja_ nägema	670	T	0,051	vastu_ võtma	379	T	0,068
tagasi_ lükkama	150	T	0,049	kokku_ leppima	183	T	0,051	kokku_ leppima	159	T	0,067
ära_ kasutama	265	T	0,049	ringi_ liikuma	97	F	0,050	peal_ kosima	1	F	0,067
sisse_ seadma	74	T	0,048	kokku_ puutuma	174	T	0,048	peale_ suruma	24	T	0,066
sisse_ astuma	76	T	0,048	alla_ neelama	114	T	0,048	eemale_ peletama	7	F	0,065
tagant_ tõukama	5	T	0,048	eemale_ tõukama	41	T	0,048	välja_ selgitama	381	T	0,065
välja_ tulema	812	T	0,048	kallale_ kargama	24	F	0,048	maha_ suruma	29	T	0,065
taga_ kiusama	10	T	0,047	tagasi_ tulema	1038	T	0,048	ringi_ sõitma	11	T	0,062
kõrvale_ heitma	26	T	0,046	vastu_ võtma	631	T	0,047	tagant_ õhutama	3	F	0,060
kõrvale_ kalduma	20	T	0,045	püsti_ ajama	175	F	0,047	esile_ kerkima	97	F	0,060
ringi_ liikuma	55	F	0,045	ligi_ pääsema	58	T	0,046	külge_ valetama	1	F	0,059
maha_ jätma	130	T	0,044	ette_ lugema	190	T	0,045	üles_ kutsuma	44	T	0,059
kokku_ hoidma	191	T	0,044	eemale_ lükkama	52	F	0,045	välja_ kujunema	344	T	0,058
tagant_ piitsutama	3	T	0,043	kinni_ panema	396	T	0,045	edasi_ lükkama	67	T	0,057
tagant_ torkima	3	T	0,043	taga_ igatsema	22	T	0,045	ümber_ kujundama	33	T	0,056
üle_ vaatama	163	T	0,043	maha_ lööma	191	T	0,044	üles_ kerkima	41	T	0,055
kallale_ kargama	4	F	0,042	ringi_ kõndima	84	T	0,043	välja_ pakkuma	312	T	0,053
ette_ võtma	393	T	0,042	kaasa_ võtma	562	F	0,042	tagant_ kannustama	1	F	0,053
ette_ heitma	135	T	0,040	ära_ võtma	562	T	0,042	üle_ kandma	112	T	0,052

50 kõrgeima MS-i väärtusega ajakirjandustekstidest tuvastatud ühendi seas on 7 ühendit (nt *eemale peletama*, *eemale tõrjuma*, *alla kukkuma*), mis ei kuulu EKSS-i loendisse. Ilukirjandustekstidest tuvastatud ühendite hulgas on seesuguseid 16 (nt *püsti tõusma*, *otsa vaatama*, *eemale lükkama*) ning teaduskirjanduses 12 (nt *külge liimima, edasi arendama*, *esile kerkima).* Seega saab ka MS kõige edukamalt hakkama ajakirjandustekstidest ühendverbide tuvastamisega, kõige halvemini aga

ilukirjandustekstidest. Igas tekstiklassis sai kõrgeid MS-i väärtuseid nii kõrge (nt *välja tulema, tagasi tulema, läbi viima*) kui ka madala sagedusega (nt *külge kleepima, tagant torkima, tagant tõukama*) õiged ühendverbid.

MS-i saab pidada ühendverbide tuvastamisel kõikidest tekstiklassidest paremaks hii-ruut-statistikust ja MI-st, kuid halvemaks t-skoorist ja log-tõepära funktsioonist.

3.1.2.6. Lihtsa sagedusloendi tulemused Tasakaalus korpuse põhjal

Tabel 9 esitab kahanevalt 50 kõige sagedasemat ühendit aja-, ilu- ja teaduskirjandustekstides.

Tabel 9. **50 sagedasemat ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb.

ajakirjandus			**ilukirjandus**			**teaduskirjandus**		
adverb_verb	**sag**	**yv**	**adverb_verb**	**sag**	**yv**	**adverb_verb**	**sag**	**yv**
välja_olema	1592	F	ära_olema	2154	F	välja_olema	1794	F
üle_olema	1022	F	välja_olema	1590	F	ette_nägema	843	T
ära_olema	1003	F	välja_tulema	1184	T	välja_tooma	817	T
kokku_olema	985	F	tagasi_tulema	1038	T	kaasa_tooma	760	T
välja_tulema	812	T	otsa_vaatama	1034	F	läbi_viima	732	T
ette_olema	802	F	läbi_olema	975	F	välja_töötama	682	T
vastu_võtma	793	T	üle_olema	907	F	üle_olema	645	F
ette_nägema	780	T	ära_tundma	838	T	ette_olema	607	F
läbi_olema	660	F	ette_olema	805	F	esile_tooma	573	T
valmis_olema	614	F	ette_kujutama	804	T	kokku_olema	540	F
välja_andma	557	T	maha_olema	762	F	ära_olema	510	F
tagasi_olema	506	F	tagasi_olema	758	F	välja_andma	501	T
tagasi_tulema	420	T	lahti_olema	738	F	läbi_olema	472	F
kinni_pidama	412	T	täis_olema	710	F	kaasa_aitama	388	T
kaasa_tooma	402	T	ära_minema	698	T	välja_selgitama	381	T
ette_võtma	393	T	kokku_olema	692	F	vastu_võtma	379	T
vastu_olema	387	F	välja_nägema	670	T	välja_kujunema	344	T
ära_tegema	387	T	kinni_olema	647	F	esile_kutsuma	316	T
kokku_saama	384	T	vastu_võtma	631	T	välja_pakkuma	312	T
üles_olema	374	F	üles_olema	575	F	esile_tõstma	310	T
maha_olema	360	F	lahti_tegema	568	T	esile_olema	305	F
alla_kirjutama	352	T	kaasa_võtma	562	F	üle_minema	283	T
välja_nägema	343	T	ära_võtma	562	T	välja_arvama	274	T

välja_kuulutama	338	T	kinni_hoidma	558	T	välja_võima	272	F
ligi_olema	335	F	valmis_olema	548	F	välja_tulema	261	T
vastu_pidama	329	T	sisse_olema	524	F	kaasa_olema	258	F
kinni_olema	325	F	koos_olema	513	F	välja_jääma	235	T
kokku_leppima	322	T	kinni_võtma	507	T	ära_kasutama	228	T
üle_andma	321	T	tagasi_minema	492	T	ära_tundma	228	T
ette_kujutama	320	T	edasi_olema	487	F	kokku_langema	225	T
ära_võtma	319	T	peale_olema	482	F	edasi_olema	224	F
tagasi_saama	317	T	välja_minema	461	T	üles_olema	221	F
välja_tooma	307	T	ette_võtma	457	T	esile_tulema	211	F
täis_olema	305	F	kokku_saama	449	T	ära_tooma	210	T
sisse_olema	304	F	püsti_tõusma	437	F	alla_olema	209	F
läbi_viima	302	T	kaasa_olema	429	F	valmis_olema	193	F
edasi_olema	297	F	lahti_saama	400	T	ette_valmistama	189	T
kaasa_olema	294	F	kinni_panema	396	T	kokku_võtma	187	T
alla_olema	286	F	vastu_pidama	390	T	edasi_andma	183	T
ette_valmistama	286	T	välja_mõtlema	388	T	kokku_puutuma	183	T
alla_jääma	282	T	edasi_minema	387	T	välja_saama	175	T
välja_saama	273	T	ära_tulema	387	T	ette_tulema	166	T
maha_võtma	272	T	välja_tegema	386	T	välja_jätma	164	T
kaasa_võtma	267	F	sisse_astuma	385	T	kokku_leppima	159	T
ära_kasutama	265	T	ilma_olema	379	F	ette_heitma	155	T
lahti_olema	265	F	maha_jätma	376	T	sisse_olema	155	F
välja_pakkuma	263	T	ringi_vaatama	376	T	ära_hoidma	152	T
kaasa_aitama	255	T	ära_tegema	365	T	üle_võtma	151	T
maha_müüma	255	T	vastu_olema	358	F	kaasa_võima	150	F
peale_olema	254	F	ära_pidama	351	T	alla_jääma	140	T

Ajakirjandustekstide 50-st sagedasemast ühendist ei kuulu 21 EKSS-i loendisse, ilukirjandustekstides on see arv 23 ja teaduskirjanduses 16. Järelikult saab lihtne sagedusloend kõige paremini hakkama teaduskirjandusest, kõige halvemini ilukirjandusest ühendverbide tuvastamisega. Sagedaste ühendite seas on palju *olema*-ühendeid (nt *välja olema*, *ära olema*, *üle olema*), mida EKSS-i loend ei sisalda. Kuna *olema* on korpuses kõige sagedasem verb, siis on ka n-ö valede ühendverbide osakaal suur. Kõige sagedasem õige ühendverb ajakirjanduses ja ilukirjanduses on *välja tulema*, teaduskirjanduses aga *ette nägema*.

Lihtsat sagedusloendit saab pidada ühendverbide tuvastamisel kõikidest tekstiklassidest võrdväärseks hii-ruut-statistikuga.

3.1.2.7. Meetodite täpsused ja saagised Tasakaalus korpuse põhjal

Eespool esitatud tabelitest võib näha, et statistiku kõrgeima väärtuse saanud 50 kandidaatpaari seas saab kõige paremini kõikidest tekstiliikidest ühendverbide tuvastamisega hakkama t-skoor, seejärel log-tõepära funktsioon ja MS. Pisut halvemad tulemused on lihtsal sagedusloendil ja hii-ruut-statistikul, halvim tulemus on MI-l.

Selleks et teada saada, kas tehtud oletused peavad paika ja kas tulemused muutuvad, kui võrdlusesse kaasata rohkem kandidaatpaare, saab võrrelda mõõdikute täpsuseid ja saagiseid. Tabel 10 esitabki olenevalt tekstiklassist meetodite täpsused ja saagised ühendverbide tuvastamisel aja-, ilu ja teaduskirjandusest, kui vaadeldavate kandidaatpaaride arv (*n*) on 50, 1000 ja 2000.

Tabel 10. **Sümmeetriliste statistikute täpsused ja saagised ühendverbide tuvastamisel aja-, ilu- ja teaduskirjandustekstidest.**

Lühendid: stat – statistik; t-sk – t-skoor, hii – hii-ruut-statistik, log – log-tõepära funktsioon, sag – lihtne sagedusloend.

stat	**meetod**	**ajakirjandus**			**ilukirjandus**			**teaduskirjandus**		
		n=50	n=1000	n=2000	n=50	n=1000	n=2000	n=50	n=1000	n=2000
t-sk	täpsus	96,0%	62,6%	42,2%	90,0%	60,3%	44,8%	92,0%	51,4%	33,1%
	saagis	2,8%	36,0%	48,6%	2,6%	34,7%	51,6%	2,6%	29,6%	38,2%
MI	täpsus	12,0%	11,9%	14,1%	4,0%	7,7%	8,1%	10,0%	17,7%	17,9%
	saagis	0,3%	6,9%	16,2%	0,1%	4,4%	9,3%	0,3%	10,2%	20,7%
hii	täpsus	82,0%	40,5%	31,6%	56,0%	38,2%	30,4%	64,0%	34,4%	27,6%
	saagis	2,4%	23,3%	36,4%	1,6%	22,0%	35,0%	1,8%	19,8%	31,8%
log	täpsus	94,0%	60,4%	38,8%	82,0%	58,1%	41,6%	90,0%	47,1%	30,2%
	saagis	2,7%	34,8%	44,7%	2,4%	33,4%	47,9%	2,6%	27,1%	34,8%
MS	täpsus	86,0%	51,0%	35,9%	68,0%	49,7%	37,1%	76,0%	38,9%	26,3%
	saagis	2,5%	29,4%	41,3%	2,0%	28,6%	42,7%	2,2%	22,4%	30,3%
sag	täpsus	58,0%	56,9%	41,6%	54,0%	59,0%	46,0%	68,0%	46,0%	30,8%
	saagis	1,7%	32,8%	48,9%	1,6%	34,0%	52,9%	2,0%	26,5%	35,5%

Tabel 10 kinnitab eespool öeldut, et t-skoor ja log-tõepära funktsioon on 50 kõrgeima statistiku väärtusega kandidaatpaari hulgas parimad meetodid ühendverbide tuvastamiseks nii aja-, ilu- kui ka teaduskirjandusest. Aja- ja ilukirjandustekstidest

ühendverbide tuvastamisel järgnevad neile täpsuse ja saagise poolest MS ja hii-ruut-statistik, kuid ajakirjandustekstidest tuvastavad nimetatud mõõdikud kõrgema täpsusega ühendverbe kui ilukirjandustekstidest. Mõlemas tekstiklassis on halvimast mõõdikust MI-st parem ka lihtne sagedusloend, millel ilukirjandustekstide korral on sarnased tulemused hii-ruut-statistikuga. Ühendverbide tuvastamisel teaduskirjandusest järgnevad t-skoorile ja log-tõepära funktsioonile MS ja lihtne sagedusloend. Viimasega võrreldes pisut kehvemad tulemused on hii-ruut-statistikul ning halvima täpsuse ja saagisega on MI. Seega 50 kandidaatpaari seas tuvastavad mõõdikud ühendverbe kõrgeima täpsuse ja saagisega ajakirjandustekstidest, madalaima täpsuse ja saagisega ilukirjandustekstidest. Selgub, et teaduskirjanduse mõõdikute paremusjärjestus on teiste tekstiklasside omadest mõnevõrra erinev: kui teistes tekstiklassides on hii-ruut-statistik lihtsast sagedusloendist kõrgema täpsuse ja saagisega, siis teaduskirjandusest ühendverbide tuvastamisel on lihtne sagedusloend hii-ruut-statistikust parem. Kokkuvõtlikult võib öelda, et 50 kandidaatpaari hulgas on kõikide tekstiliikide korral mõõdikute paremusjärjestus sarnane.

Kui aga vaadelda, kas ja kuidas muutuvad mõõdikute tulemused statistiku alusel järjestatud kandidaatpaaride arvu kasvamisega, siis on ajakirjandustekstidest ühendverbide tuvastamisel näha, et t-skoor on ka 2000 kandidaatpaari lõikes kõige kõrgema täpsusega mõõdik, samas on lihtsa sagedusloendi saagis natuke kõrgem kui t-skoori oma. Järelikult võib lihtsat sagedusloendit pidada sama edukaks kui t-skoori. MS-i täpsus on märgatavalt langenud kandidaatpaaride arvu suurenemisega, kuid see jääb kõrgemaks kui hii-ruut-statistiku täpsus. MI on ka 2000 kandidaatpaari lõikes teistest väiksema täpsuse ja saagisega, samas pole MI täpsus ja saagis nii palju muutunud kui teiste mõõdikute omad.

Ilukirjandustekstidest genereeritud 2000 kandidaatpaari seas on kõige edukam lihtne sagedusloend, mille täpsus on t-skoori täpsusest kõrgem 1,2%, saagis 1,3%. Log-tõepära funktsiooni täpsus ja saagis on kõrgemad kui MS-i, hii-ruut-statistiku ja MI täpsused ja saagised. Statistiku alusel järjestatud kandidaatpaaride arvu kasvamine suurendab MI täpsuseid, kuid MI tulemused jäävad teiste mõõdikute tulemustega võrreldes märgatavalt madalamaks. Järelikult vaadeldavate kandidaatpaaride arvu suurenemisega muutub statistikute paremusjärjestus ilukirjandusest ühendverbide tuvastamisel.

Teaduskirjandustekstidest genereeritud 2000 kandidaatpaari hulgas esineb kõige paremini t-skoor, kuigi selle täpsus ja saagis on võrreldes teiste tekstiklassidega väiksem. Lihtsa sagedusloendi ja log-tõepära funktsiooni täpsused ja saagised on 2000 kandidaatpaari seas peaaegu võrdsed, kandidaatpaaride arvu suurenemisega on märgatavalt kahanenud MS-i täpsus, mis on madalam kui hii-ruut-statistiku täpsus. MI täpsus on küll teistest väiksem, kuid on ainsana kandidaatpaaride arvu suurenedes kasvanud. MI saagis on teaduskirjanduse 2000 kandidaatpaari seas kõrgem kui teiste tekstiliikide sama arvu kandidaatpaaride hulgas.

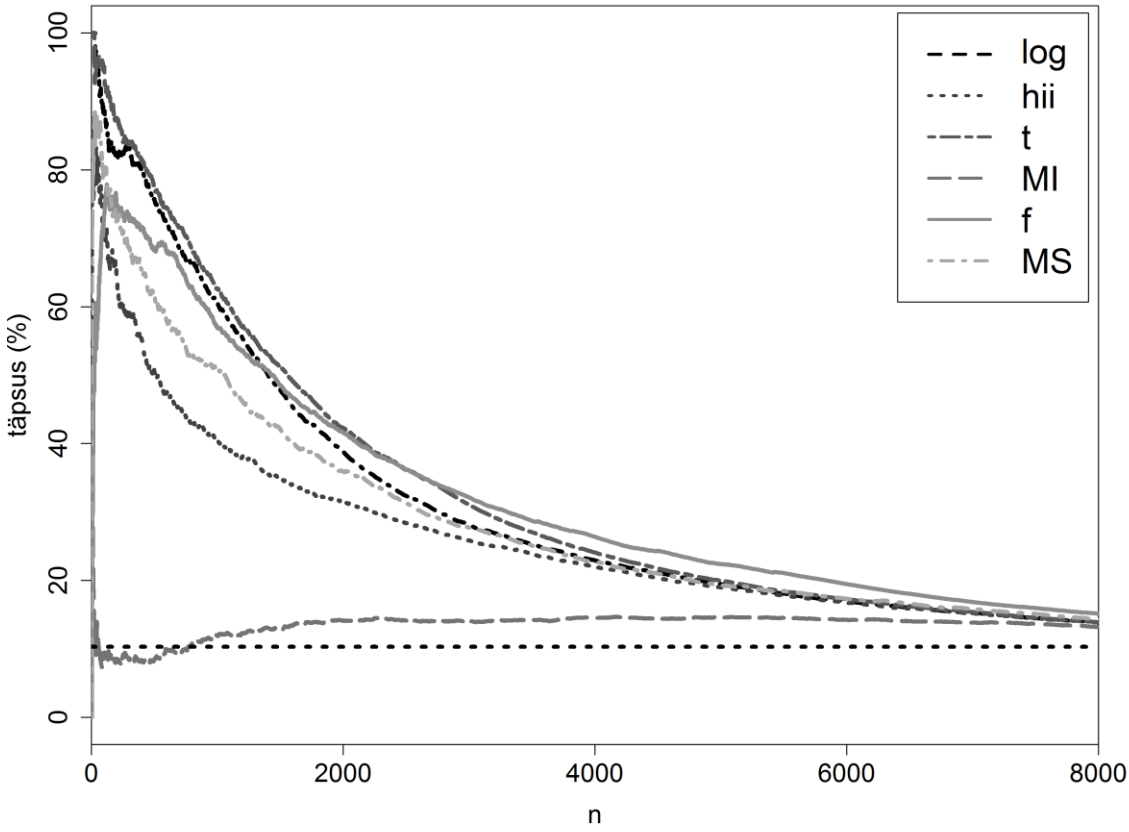
Kokkuvõtlikult saab öelda, et statistiku alusel järjestatud kandidaatpaaride arvu kasvamine muudab mõõdikute tulemusi ja paremusjärjestust. Kui t-skoor on nii 50 kandidaatpaari seas kõige edukam kõikidest tekstiklassidest ühendverbide tuvastamisel, siis 2000 kandidaatpaari hulgas on ilukirjandusest sama ülesande lahendamisel edukaim lihtne sagedusloend. Ajakirjandustekstide 2000 kandidaatpaari seas on t-skoori ja lihtsa sagedusloendi tulemused võrdsed. Heade tulemustega on ka log-tõepära funktsioon, mis kõikide tekstiliikide korral on t-skoori ja lihtsa sagedusloendi järel paremuselt kolmas mõõdik. Kui kõikide teiste statistikute täpsused langevad kandidaatpaari arvu suurendamisega, siis MI täpsus paraneb. Üldiselt on mõõdikute täpsused ja saagised kõrgeimad ilukirjandustekstidest, madalaimad teaduskirjandustekstides ühendverbide tuvastamisel. Vaid MI täpsus ja saagis on kõrgeimad teaduskirjanduse puhul.

3.1.2.8. Meetodite täpsuse kõverad Tasakaalus korpuse põhjal

Seda, kuidas mõõdikute tulemused erinevatest tekstiklassides muutuvad veel suurema hulga kandidaatpaaride hulgas, illustreerivad joonised 1, 2 ja 3, kus on esitatud statistikute ja lihtsa sageduse täpsuse kõverad kuni 8000 kandidaatpaarini.

Joonisel 1 on esitatud sümmeetriliste statistikute ja lihtsa sagedusloendi täpsuse kõverad Tasakaalus korpuse ajakirjandustekstidest ühendverbide tuvastamisel. Joonise x-telg tähistab kandidaatpaaride arvu ja y-telg meetodi täpsust. Horisontaalne punktiirjoon joonisel tähistab algtaseme täpsust, mis on saadud korpusest leitud õigete ühendverbide arvu (1351) jagamisel kõikide saadud kandidaatpaaride arvuga (13141) järgmiselt:

*algtaseme täpsus = (1351/13141)*100 ≈ 10,3%*



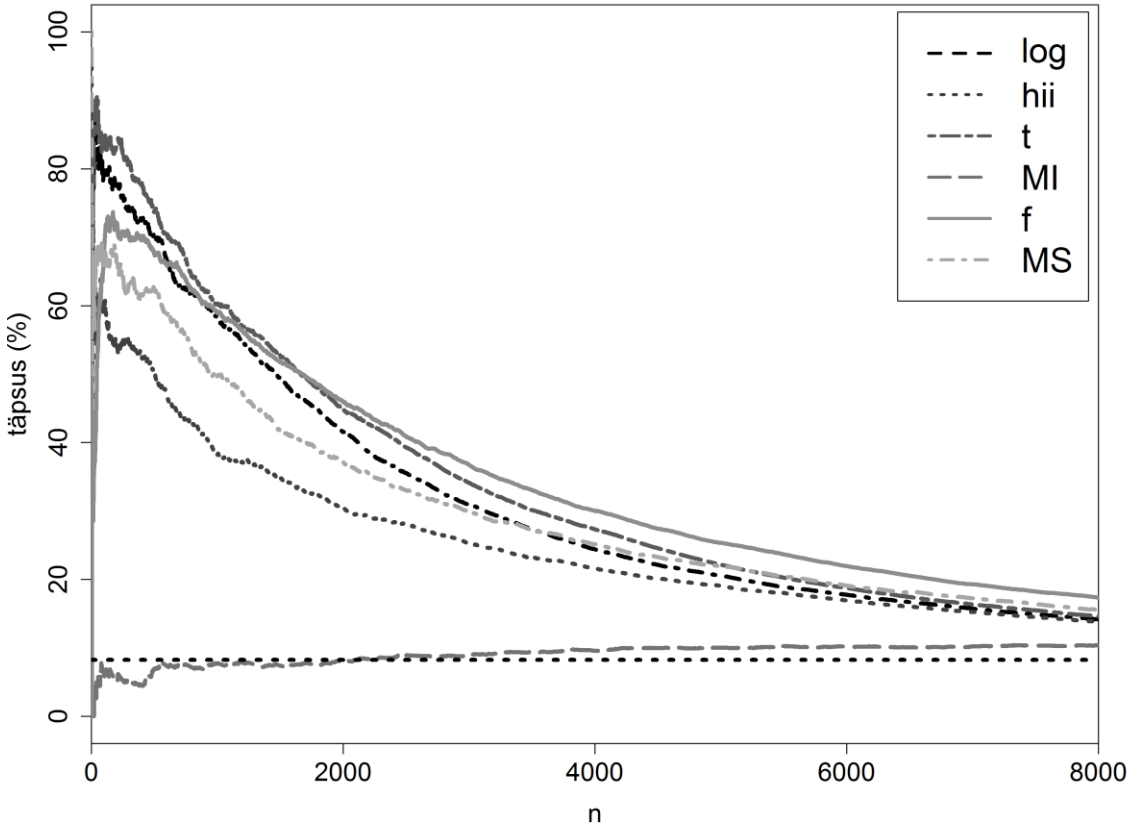
Joonis 1. **Sümmeetriliste statistikute ja sageduse täpsuse kõverad ühendverbide tuvastamisel ajakirjandustekstidest.**

Selgub, et ajakirjandustekstides on kuni 2000 kandidaatpaari hulgas t-skoor kõige tulemuslikum, kuid lihtsa sagedusloendi täpsus on peaaegu samaväärne. Alates 2000 kandidaatpaarist on aga lihtsa sagedusloendi täpsus kõige kõrgem. Väiksema arvu kandidaatpaaride seas hästi esinenud log-tõepära funktsiooni, MS-i ja hii-ruut-statistiku täpsused on alates 4000 kandidaatpaarist sarnased. MI täpsus, mis alates 1000 kandidaatpaarist kasvab, muutub teiste mõõdikutega sarnaseks 8000 kandidaatpaari hulgas. Siiski on kõik vaadeldud meetodid heade tulemustega ajakirjandustekstidest ühendverbide tuvastamisel. Peab arvestama aga sellega, et MI muutub tulemuslikuks alates 1000 kandidaatpaarist ja selle täpsus paraneb kandidaatpaaride arvu kasvamisega.

Joonis 2 illustreerib statistikute ja lihtsa sagedusloendi täpsuse kõveraid ühendverbide tuvastamisel ilukirjandustekstidest. Joonise x-telg tähistab

kandidaatpaaride arvu ja y-telg meetodi täpsust. Horisontaalne punktiirjoon joonisel tähistab algtaseme täpsust, mis on saadud korpusest leitud õigete ühendverbide arvu (1519) jagamisel kõikide genereeritud kandidaatpaaride arvuga (18459) järgmiselt:

*algtaseme täpsus = (1519/18459)*100 ≈ 8,2%*



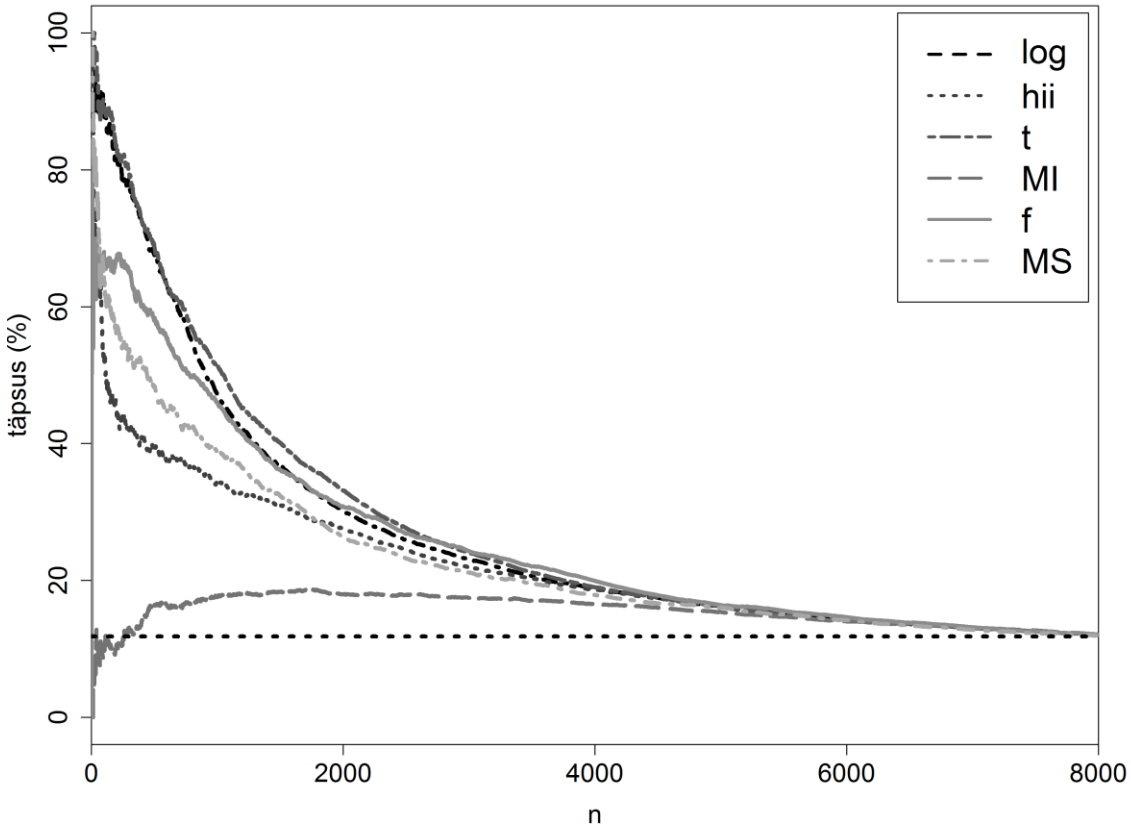
Joonis 2. **Sümmeetriliste statistikute ja sageduse täpsuse kõverad ühendverbide tuvastamisel ilukirjandustekstidest.**

Ilukirjandusetekstidest tuvastab ühendverbe esimese 1000 kandidaatpaari hulgast kõige paremini t-skoor, kuid joonis 2 kinnitab eelmises peatükis tehtud järeldust, et 2000 kandidaatpaari seas on lihtne sagedus efektiivsem kui t-skoor. Kandidaatpaaride arvu suurenemisega võrdsustub ka mõõdikute täpsus, kuid lisaks lihtsale sagedusloendile ja t-skoorile on ülesande lahendamisel tulemuslikud ka log-tõepära funktsioon ja MS ning pisut halvemate tulemustega ka hii-ruut statistik. MI esineb taaskord kõige halvemini –

statistiku alusel järjestatud kandidaatpaaride arvu suurenemisel MI täpsus küll paraneb, kuid jääb algtaseme täpsuse joonega peaaegu samale tasemele ja MI-d võib kasulikuks lugeda alles siis, kui arvesse võetakse 3000 või rohkem kandidaatpaari. Seega on MI ühendverbide tuvastamisel ajakirjandustekstidest efektiivsem kui ilukirjandustekstidest.

Joonis 3 esitab sümmeetriliste statistikute ja lihtsa sagedusloendi täpsuse kõverad teaduskirjandusest ühendverbide tuvastamisel. Joonise x-telg tähistab kandidaatpaaride arvu ja y-telg meetodi täpsust. Horisontaalne punktiirjoon joonisel tähistab algtaseme täpsust, mis on saadud korpusest leitud õigete ühendverbide arvu (972) jagamisel kõikide saadud kandidaatpaaride arvuga (8218) järgmiselt:

*algtaseme täpsus = (972/8218)*100 ≈ 11,8%*



Joonis 3. **Sümmeetriliste statistikute ja sageduse täpsuse kõverad ühendverbide tuvastamisel teaduskirjandustekstidest.**

Nagu ka eespool selgus on log-tõepära funktsiooni tulemused teaduskirjanduse 50 kandidaatpaari hulgas sarnased t-skoori tulemustele. Kuigi t-skoor on parimaid tulemusi andev meetod nii 50 kui ka 2000 kandidaatpaari seas, on alates 3000 kandidaatpaarist parim lihtne sagedusloend, mis 2000 kandidaatpaari juures on natuke parem log-tõepära funktsioonist. Kui 2000 kandidaatpaari hulgas on hii-ruut-statistiku, MS-i ja MI tulemused halvemad kui teiste omad, siis 4000 kandidaatpaari seas on meetodite täpsused üsna sarnased. Alates 100 kandidaatpaarist tõuseb MI väärtus pidevalt ja kui kombineerida mõõdikut teistega, võib seda pidada tulemuslikuks ühendverbide tuvastamisel teaduskirjanduse korpusest.

3.1.2.9. Kokkuvõte sümmeetriliste mõõdikute ja lihtsa sagedusloendi tulemustest Tasakaalus korpuse põhjal

Sümmeetrilised statistikud töötavad võrreldud tekstiklassidest ühendverbide tuvastamisel sarnaselt, kuid mitte täpselt ühtemoodi. 50 kandidaatpaari hulgas on kõige efektiivsem kõikide tekstiklasside korral t-skoor, kuid kandidaatpaaride arvu suurenemisega tulemused muutuvad. Kui aja- ja teaduskirjandustekstide 2000 kandidaatpaari seas on t-skoor endiselt kõrgeima täpsusega, siis ilukirjandusest ühendverbide tuvastamisel on kõige efektiivsem lihtne sagedusloend. 50 kandidaatpaari hulgas on kõikides tekstiklassides mõõdikute järjestus sarnane: parim on t-skoor, sellele järgnevad log-tõepära, MS, hii-ruut-statistik, lihtne sagedusloend ja MI. Vaid teaduskirjanduse korral on lihtne sagedusloend hii-ruut-statistikust parem. 2000 kandidaatpaari seas on log-tõepära funktsioon paremusjärjestuses kolmas. Kui aja- ja ilukirjandustekstidest ühendverbide tuvastamisel järgneb eelmistele neljandana MS, viiendana hii-ruut-statistik ja viimasena MI, siis teaduskirjandusest tuvastab hii-ruut-statistik ühendverbe paremini kui MS. MI täpsus on küll madal iga tekstiklassi korral, kuid MI on ainus mõõdik, mille täpsus kandidaatpaaride arvu suurenedes paraneb. Kõikide tekstiklasside kohta saab öelda, et alates 4000 kandidaatpaarist on lihtne sagedusloend efektiivseim mõõdik.

Peaaegu kõiki vaadeldud meetodeid – t-skoori, hii-ruut-statistikut, log-tõepära funktsiooni, MS-i ja lihtsat sagedusloendit – võib pidada tulemuslikeks ühendverbide tuvastamisel erinevaid tekstiliike sisaldavatest korpustest. MI tulemuslikkus on seda parem, mida rohkem kõikidest genereeritud kandidaatpaaridest vaatluse all on.

3.1.3. Asümmeetriliste mõõdikute tulemused Tasakaalus korpuse põhjal

See peatükk kirjeldab asümmeetriliste mõõdikute tulemusi ühendverbide tuvastamisel Tasakaalus korpusesse kuuluvatest tekstiklassidest. Eesmärk on võrrelda asümmeetriliste statistikute tulemusi tekstiklasside lõikes.

3.1.3.1. Tingliku tõenäosuse tulemused Tasakaalus korpuse põhjal

Tabelid 11 ja 12 esitavad kahanevalt 50 kõrgeima tingliku tõenäosuse väärtuse saanud ühendit aja-, ilu- ja teaduskirjandustekstides. Tabelis 11 esitatud tingliku tõenäosuse väärtused väljendavad, kui suur on verbi esinemise tõenäosus, kui samas osalauses esineb ühendisse kuuluv adverb. Tabelis 12 esitatud tingliku tõenäosuse väärtused väljendavad, kui suure tõenäosusega esineb ühendisse kuuluv adverb, kui samas osalauses esineb ühendisse kuuluv verb.

Tabel 11. **50 kõrgeima tingliku tõenäosuse CP(verb|adverb) väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; sag – ühendi koosesinemissagedus samas osalauses, CP(2|1) – CP(verb|adverb) väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**CP(2|1)**	**adverb_ verb**	**sag**	**yv**	**CP(2|1)**	**adverb_ verb**	**sag**	**yv**	**CP(2|1)**
pärale_ jõudma	32	T	0,941	pärale_ jõudma	98	T	0,961	kallale_ tungima	2	F	1
ülal_ pidama	75	T	0,882	kaasas_ olema	342	F	0,691	ühes_ toimuma	1	F	1
valmis_ olema	614	F	0,667	järel_ olema	129	F	0,645	ühes_ pidama	1	F	1
järel_ olema	49	F	0,653	valmis_ olema	548	F	0,626	lahku_ minema	30	T	0,769
tasa_ tegema	37	T	0,597	otsa_ vaatama	1034	F	0,606	pärale_ jõudma	6	T	0,750
kaasas_ olema	206	F	0,572	lahku_ minema	93	T	0,592	järel_ olema	21	F	0,750
vahele_ jääma	160	T	0,557	peal_ olema	98	F	0,551	täis_ olema	43	F	0,717
üleval_ olema	84	F	0,538	täis_ olema	710	F	0,516	valmis_ olema	193	F	0,715
lahku_ minema	57	T	0,538	üleval_ olema	160	F	0,466	pihta_ saama	9	T	0,643
täis_ olema	305	F	0,535	pihta_ saama	97	T	0,401	valla_ olema	13	F	0,591
peal_ olema	53	F	0,477	koos_ olema	513	F	0,392	vastu_ võtma	379	T	0,485

alt_ vedama	55	T	0,437	ilma_ olema	379	F	0,388	peal_ olema	5	F	0,455
pihta_ hakkama	53	T	0,434	pihta_ hakkama	91	T	0,376	tasa_ olema	9	F	0,450
taha_ jääma	11	T	0,407	kõrvalt_ vaatama	48	T	0,369	kaasas_ käima	51	T	0,447
valla_ olema	47	F	0,382	ülal_ pidama	44	T	0,361	pealt_ olema	16	F	0,444
pihta_ saama	46	T	0,377	püsti_ tõusma	437	F	0,343	kaasa_ tooma	760	T	0,443
ilma_ jääma	243	T	0,377	valla_ olema	51	F	0,340	läbi_ viima	732	T	0,436
vastu_ võtma	793	T	0,368	taga_ ajama	143	T	0,291	taga_ olema	22	F	0,431
kõrvalt_ vaatama	26	T	0,342	pärale_ olema	29	F	0,284	ette_ nägema	843	T	0,423
taga_ olema	73	F	0,340	pealt_ vaatama	104	T	0,279	tasa_ tegema	8	T	0,400
koos_ olema	248	F	0,309	taga_ olema	131	F	0,266	taha_ minema	2	F	0,400
otsa_ saama	128	T	0,293	ülal_ olema	32	F	0,262	taha_ olema	2	F	0,400
ühte_ olema	14	F	0,286	ühte_ olema	37	F	0,257	ligi_ olema	112	F	0,392
külge_ olema	25	F	0,281	läbi_ olema	975	F	0,245	peale_ olema	134	F	0,371
esile_ tõstma	136	T	0,278	esile_ kutsuma	75	T	0,240	ilma_ olema	137	F	0,368
ligi_ olema	335	F	0,274	kaasa_ võtma	562	F	0,238	otsa_ saama	18	T	0,367
üle_ olema	1022	F	0,271	sekka_ olema	14	F	0,237	kõrvalt_ olema	4	F	0,364
külge_ jääma	24	T	0,270	kallale_ tulema	48	F	0,234	kõrvale_ jätma	130	T	0,361
läbi_ olema	660	F	0,253	ümber_ olema	268	F	0,233	kinni_ pidama	95	T	0,358
kinni_ pidama	412	T	0,246	kõrvalt_ olema	30	F	0,231	esile_ tooma	573	T	0,354
peale_ olema	254	F	0,244	peale_ olema	482	F	0,230	kaasas_ olema	40	F	0,351
lahti_ olema	265	F	0,244	üle_ olema	907	F	0,229	üleval_ olema	12	F	0,343
ilma_ olema	157	F	0,243	pealt_ olema	79	F	0,212	vahele_ jääma	15	T	0,333
tasa_ olema	15	F	0,242	vahele_ jääma	92	T	0,207	külge_ olema	5	F	0,333
laiali_ olema	69	F	0,240	külge_ olema	46	F	0,204	lahti_ olema	63	F	0,317
otsa_ sõitma	105	T	0,240	lahti_ olema	738	F	0,204	üleval_ pidama	11	T	0,314
ette_ olema	802	F	0,235	tagasi_ tulema	1038	T	0,201	vahele_ jätma	14	T	0,311
taga_ ajama	50	T	0,233	vastu_ võtma	631	T	0,199	välja_ olema	1794	F	0,305

ette_ nägema	780	T	0,229	mööda_ minema	221	T	0,197	ette_ olema	607	F	0,305
ümber_ olema	177	F	0,227	ligi_ olema	108	F	0,197	alla_ olema	209	F	0,299
kokku_ olema	985	F	0,226	ära_ olema	2154	F	0,196	üle_ olema	645	F	0,298
kõrvalt_ olema	17	F	0,224	ringi_ vaatama	376	T	0,194	sisse_ olema	155	F	0,298
eemale_ jääma	53	T	0,221	kokku_ olema	692	F	0,193	üles_ olema	221	F	0,297
esile_ kutsuma	105	T	0,215	esile_ tooma	60	T	0,192	järele_ olema	46	F	0,286
sekka_ olema	9	F	0,214	ette_ olema	805	F	0,191	ülal_ olema	22	F	0,286
alla_ kirjutama	352	T	0,213	ette_ kujutama	804	T	0,191	läbi_ olema	472	F	0,281
esile_ tooma	103	T	0,211	alt_ olema	44	F	0,186	kõrvalt_ vaatama	3	T	0,273
üleval_ pidama	32	T	0,205	külge_ jääma	41	T	0,182	mööda_ minema	34	T	0,266
pealt_ nägema	30	T	0,201	ilma_ jääma	178	T	0,182	püsti_ tõusma	17	F	0,262
ülal_ olema	17	F	0,200	kaasa_ olema	429	F	0,182	kinni_ olema	69	F	0,260

Tabel 12. **50 kõrgeima tingliku tõenäosuse CP(adverb|verb) väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.**

Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; CP(1|2) – CP(adverb|verb) väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**CP(1|2)**	**adverb_ verb**	**sag**	**yv**	**CP(1|2)**	**adverb_ verb**	**sag**	**yv**	**CP(1|2)**
sekka_ hällitama	1	F	1	ülal_ ekshibeerima	1	F	1	valla_ pääsima	1	F	1
tagant_ suskima	1	F	1	laiali_ sõõrduma	5	F	1	järel_ vantsima	1	F	1
püsti_ krapsama	1	F	1	tagant_ punima	1	F	1	üleval_ semantiseeruma	1	F	1
laiali_ hoorama	1	F	1	tagant_ utjama	1	F	1	vahele_ tükkima	1	F	1
vahele_ pläristama	1	F	1	valla_ loppuma	1	F	1	alla_ kriipsutama	15	T	1
kaasas_ tilpnema	1	F	1	lahku_ translitereerima	1	F	1	taga_ lähnuma	1	F	1
mööda_ sudima	1	F	1	peal_ ruiguma	1	F	1	ühte_ kollabseeruma	1	F	1
järele_ reformeerima	1	F	1	kallale_ hassetama	1	F	1	püsti_ kohisema	1	F	1
sisse_ logima	3	T	1	alt_ kolmandama	1	F	1	lahti_ rulluma	3	T	1
peale_ tarretuma	2	F	1	alt_ muduma	1	F	1	lahti_ muukima	2	F	1

täis_ maruma	1	F	1	üleval_ poositama	1	F	1	mööda_ visioneerima	1	F	1
täis_ roojama	1	F	1	pealt_ ilgema	1	F	1	kinni_ tukastama	2	F	1
ilma_ etiketistama	1	F	1	pealt_ satuma	1	F	1	järele_ ahvima	1	F	1
ilma_ kekutama	1	F	1	kaasas_ taastunnustama	1	F	1	järele_ võsastuma	1	F	1
ringi_ floresiensima	1	F	1	ümber_ plartseldama	2	F	1	järele_ önnistuma	1	F	1
ringi_ kooberdama	1	F	1	laiali_ kluksuma	1	F	1	ringi_ meikima	1	F	1
ringi_ melutsema	1	F	1	laiali_ siia-sinama	1	F	1	ringi_ pöruma	1	F	1
ringi_ roobeldama	1	F	1	eemale_ lootsima	1	F	1	ringi_ sobrama	1	F	1
välja_ joonistuma	13	F	1	valmis_ schoolima	1	F	1	ringi_ vurama	1	F	1
ümber_ grupeeruma	1	F	1	ilma_ jõutama	1	F	1	lahti_ kasukkama	1	F	1
koos_ armeerima	1	F	1	mööda_ ehituma	1	F	1	lahti_ korkima	1	T	1
koos_ kudrutama	1	F	1	mööda_ neli-viima	1	F	1	ümber_ kiskuma	2	F	1
koos_ kypsetama	1	F	1	mööda_ vuhkima	1	F	1	sisse_ kasseerima	2	F	1
valmis_ jäigendama	1	F	1	ümber_ defineeruma	1	F	1	kinni_ valendama	1	F	1
valmis_ plökerdama	1	F	1	ümber_ parantsatama	1	F	1	kõrvale_ põiklema	1	F	1
valmis_ sheigima	1	F	1	ümber_ plumpsuma	1	F	1	peale_ illotuma	1	F	1
peale_ härduma	1	F	1	ümber_ profileerima	1	F	1	ilma_ indemnitama	1	F	1
lahti_ kirjutatama	1	F	1	püsti_ ehituma	1	F	1	ilma_ sandistama	1	F	1
lahti_ magneetima	1	F	1	püsti_ kräuksatama	1	F	1	maha_ kopuleeruma	1	F	1
lahti_ raspeldama	1	F	1	püsti_ looberdama	1	F	1	maha_ kõmpima	1	T	1
lahti_ suutduma	1	F	1	püsti_ virtsama	1	F	1	maha_ materdama	1	F	1
ligi_ taasinvesteerima	1	F	1	täis_ köettuma	1	F	1	maha_ prassima	1	F	1
sisse_ canariensima	1	F	1	täis_ litsitama	1	F	1	maha_ tolgotama	1	F	1
sisse_ desifintseerima	1	F	1	täis_ plätsutama	1	F	1	maha_ võtima	1	F	1
sisse_ freesima	1	F	1	vastu_ kõpsima	2	F	1	ümber_ portima	1	F	1
sisse_ lõõskama	1	F	1	sisse_ tsuskama	2	F	1	ümber_ rahvastuma	1	F	1
sisse_ melutsema	1	F	1	otsa_ krüptima	1	F	1	ümber_ suupärastama	1	F	1
sisse_ susserdama	1	F	1	otsa_ tõsima	1	F	1	ümber_ taasäratama	1	F	1

sisse_ talluma	1	F	1	järele_ kösima	1	F	1	sisse_ aerutama	1	F	1
sisse_ võmmutama	1	F	1	järele_ lausima	1	F	1	sisse_ annoteerima	1	F	1
alla_ aetama	1	F	1	järele_ lõõgastama	1	F	1	sisse_ kreemitama	1	F	1
alla_ laanima	1	F	1	järele_ tšekkama	1	F	1	sisse_ meldima	1	F	1
kinni_ graundima	1	F	1	kokku_ kõlksahtama	2	F	1	üle_ kavaldama	4	T	1
kinni_ klammerdama	1	F	1	edasi_ jämmima	2	F	1	ette_ atakeerima	3	F	1
kinni_ mürtsatama	1	F	1	ringi_ ehituma	1	F	1	ära_ atakeerima	3	F	1
ära_ vussima	3	F	1	ringi_ helklema	1	F	1	alla_ kantima	1	F	1
üles_ lebasklema	1	F	1	ringi_ lidisema	1	F	1	alla_ sähvatama	1	F	1
üles_ manageerima	1	F	1	ringi_ lidristama	1	F	1	alla_ väärtustatama	1	F	1
üles_ nõmetsema	1	F	1	ringi_ lööberdama	1	F	1	üles_ miesima	1	F	1
vastu_ margistama	1	F	1	ringi_ põrklema	1	F	1	üles_ taaskiirguma	1	F	1

Ajakirjandustekstides said kõige tugevama CP(verb|adverb) väärtuse õiged ühendverbid *pärale jõudma* ja *ülal pidama*. Tabelis 11 esitatud tinglik tõenäosus väljendab, kui tugevalt sõltub verbi esinemine adverbist ja seetõttu saab öelda, et kui osalauses esineb sõna *pärale*, siis esineb seal suure tõenäosusega verb *jõudma.* Samamoodi on ühendiga *ülal pidama*. Nimetatud ühendid on ainsad ühendverbid, mis on EKSS-i nimistus adverbidega *pärale* ja *ülal* moodustatud. Ilukirjandustekstides on kõrge tinglikku tõenäosuse väärtuse saanud mitmed *olema*-ühendid. Kui aja- ja ilukirjandustekstides pole ükski ühend saanud väärtuseks 1, siis teaduskirjanduses on neid kolm: *kallale tungima*, *ühes toimuma*, *ühes pidama*. Nende kolme ühendi esinemise kohta võib öelda, et kui osalauses esineb adverb, siis esineb alati samas osalauses ka temaga ühendi moodustav verb. Kui vaadata teaduskirjanduse andmestikku, siis selgub, et *ühes* on vaid korra saanud selles korpuses adverbi märgendi: lauses *[[Varsti täiendati nimetatud reeglit kohaühtsuse nõudega,] [mille järgi näidendi tegevus peab toimuma ühes- ja sellessamas paigas.]]*. Adverb *ühes* on samas osalauses verbidega *pidama* ja *toimuma*, mistõttu esinevadki *ühes pidama* ja *ühes toimuma* korpuses korra. Sõna *kallale* esineb teaduskirjanduse korpuses kaks korda, mõlemal korral on ta samas osalauses verbiga *tungima*.

Kõige paremini tuvastab CP(verb|adverb) ajakirjandustekstidest õigeid ühendverbe – 50 kõrgeima väärtusega ühendi seas on neid 25. Nii ilu- kui ka teaduskirjandustekstidest tuvastab CP(verb|adverb) õigeid ühendverbe 18.

Kui vaadelda tinglikku tõenäosuse CP(adverb|verb) tulemusi, kus adverbi esinemine sõltub verbi esinemisest, siis näeb, et kõik 50 kõrgeima skooriga ühendit on saanud väärtuseks 1. See tähendab, et nende ühendite komponentide vahel kehtib reegel, et kui osalauses esineb ühendis olev verb, siis alati esineb seal ka ühendisse kuuluv adverb. Nagu tabelist 12 näha on, siis tegemist on harva esinevate ühenditega, mille verbiline osa pole väga sage. Nii tuvastabki tinglik tõenäosus CP(adverb|verb) ajakirjandustekstidest vaid ühe õige ühendverbi (*sisse logima*), ilukirjandustekstidest aga mitte ühtegi. Teaduskirjandusest ühendverbide tuvastamisel on aga CP(adverb|verb) edukam ja tuvastab viis õiget ühendverbi (*alla kriipsutama*, *lahti rulluma*, *lahti korkima*, *maha kõmpima*, *üle kavaldama*).

3.1.3.2. ΔP tulemused Tasakaalus korpuse põhjal

Tabelid 13 ja 14 esitavad kahanevalt 50 kõrgeima ΔP väärtuse saanud ühendit aja-, ilu- ja teaduskirjandustekstides. Tabelis 13 esitatud ΔP väärtused väljendavad, kui tugevalt sõltub verbi esinemine adverbi esinemisest samas osalauses, tabelis 14 esitatud ΔP väärtused aga seda, kui tugevalt sõltub ühendisse kuuluva adverbi esinemine verbi esinemisest samas osalauses.

Tabel 13. **50 kõrgeima ΔP(verb|adverb) väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.** Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; ΔP(2|1) – ΔP(verb|adverb) väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**ΔP(2|1)**	**adverb_ verb**	**sag**	**yv**	**ΔP(2|1)**	**adverb_ verb**	**sag**	**yv**	**ΔP(2|1)**
pärale_ jõudma	32	T	0,934	pärale_ jõudma	98	T	0,954	kallale_ tungima	2	F	1
ülal_ pidama	75	T	0,853	otsa_ vaatama	1034	F	0,594	ühes_ toimuma	1	F	0,993
tasa_ tegema	37	T	0,572	lahku_ minema	93	T	0,569	ühes_ pidama	1	F	0,978
vahele_ jääma	160	T	0,543	kaasas_ olema	342	F	0,416	lahku_ minema	30	T	0,766
lahku_ minema	57	T	0,522	järel_ olema	129	F	0,370	pärale_ jõudma	6	T	0,747

alt_ vedama	55	T	0,435	pihta_ saama	97	T	0,368	pihta_ saama	9	T	0,617
pihta_ hakkama	53	T	0,423	pihta_ hakkama	91	T	0,362	vastu_ võtma	379	T	0,477
taha_ jääma	11	T	0,392	kõrvalt_ vaatama	48	T	0,357	järel_ olema	21	F	0,474
valmis_ olema	614	F	0,382	valmis_ olema	548	F	0,351	kaasas_ käima	51	T	0,446
järel_ olema	49	F	0,368	püsti_ tõusma	437	F	0,340	täis_ olema	43	F	0,440
ilma_ jääma	243	T	0,362	ülal_ pidama	44	T	0,338	valmis_ olema	193	F	0,439
vastu_ võtma	793	T	0,356	taga_ ajama	143	T	0,287	kaasa_ tooma	760	T	0,437
kõrvalt_ vaatama	26	T	0,337	peal_ olema	98	F	0,276	läbi_ viima	732	T	0,434
pihta_ saama	46	T	0,333	pealt_ vaatama	104	T	0,266	ette_ nägema	843	T	0,419
kaasas_ olema	206	F	0,287	täis_ olema	710	F	0,241	taha_ minema	2	F	0,397
esile_ tõstma	136	T	0,276	esile_ kutsuma	75	T	0,237	tasa_ tegema	8	T	0,387
külge_ jääma	24	T	0,255	kaasa_ võtma	562	F	0,224	kõrvale_ jätma	130	T	0,359
üleval_ olema	84	F	0,254	kallale_ tulema	48	F	0,210	esile_ tooma	573	T	0,347
täis_ olema	305	F	0,250	vahele_ jääma	92	T	0,193	otsa_ saama	18	T	0,342
otsa_ saama	128	T	0,249	üleval_ olema	160	F	0,192	kinni_ pidama	95	T	0,336
otsa_ sõitma	105	T	0,236	ette_ kujutama	804	T	0,190	vahele_ jääma	15	T	0,325
taga_ ajama	50	T	0,231	esile_ tooma	60	T	0,188	valla_ olema	13	F	0,315
ette_ nägema	780	T	0,221	vastu_ võtma	631	T	0,185	vahele_ jätma	14	T	0,309
kinni_ pidama	412	T	0,217	ringi_ vaatama	376	T	0,182	üleval_ pidama	11	T	0,292
esile_ kutsuma	105	T	0,212	valla_ pääsema	27	T	0,179	kõrvalt_ vaatama	3	T	0,271
alla_ kirjutama	352	T	0,208	tagasi_ tulema	1038	T	0,178	mööda_ minema	34	T	0,262
eemale_ jääma	53	T	0,206	mööda_ minema	221	T	0,174	püsti_ tõusma	17	F	0,260
esile_ tooma	103	T	0,205	külge_ jääma	41	T	0,169	taga_ ajama	13	T	0,254
pealt_ nägema	30	T	0,193	ilma_ jääma	178	T	0,169	sekka_ sattuma	1	F	0,249
peal_ olema	53	F	0,193	valla_ päästma	25	T	0,166	sekka_ rääkima	1	F	0,247
kaasas_ käima	71	T	0,190	pealt_ nägema	66	T	0,163	pärale_ viima	2	F	0,246
kõrvale_ jääma	89	T	0,184	pealt_ kuulama	57	T	0,151	sekka_ looma	1	F	0,245

kaasa_ tooma	402	T	0,180	vahele_ segama	66	T	0,147	mööda_ saama	33	T	0,232
püsti_ tõusma	51	F	0,177	alt_ vedama	35	T	0,147	kaasa_ aitama	388	T	0,224
üleval_ pidama	32	T	0,176	kallale_ tungima	30	F	0,146	sekka_ võima	1	F	0,219
pealt_ kuulama	25	T	0,167	peale_ hakkama	334	T	0,146	laiali_ saatma	16	F	0,218
kallale_ tulema	18	F	0,165	lahti_ tegema	568	T	0,135	ülal_ pidama	18	T	0,211
püsti_ panema	48	F	0,161	kinni_ hoidma	558	T	0,134	tagant_ tõukama	4	T	0,210
alla_ jääma	282	T	0,156	püsti_ ajama	175	F	0,134	pihta_ hakkama	3	T	0,209
mööda_ minema	72	T	0,151	kõrvalt_ jälgima	17	F	0,130	maha_ jääma	97	T	0,208
üleval_ hoidma	24	F	0,151	ringi_ käima	262	T	0,128	taha_ paisuma	1	F	0,200
eemale_ hoidma	36	F	0,147	ligi_ astuma	72	F	0,127	taha_ vallandama	1	F	0,200
vahele_ jätma	43	T	0,146	ühte_ sulama	18	T	0,125	taha_ kogunema	1	F	0,200
järele_ jääma	80	T	0,145	valmis_ saama	135	T	0,122	taha_ tõmbama	1	F	0,200
juurde_ tulema	120	T	0,144	koos_ olema	513	F	0,117	taha_ kalduma	1	F	0,200
eemale_ peletama	34	F	0,142	kallale_ kargama	24	F	0,117	taha_ järgima	1	F	0,199
tagant_ sõitma	10	F	0,140	koos_ elama	158	F	0,115	taha_ õnnestuma	1	F	0,199
taha_ panema	4	T	0,140	eemale_ hoidma	100	F	0,113	sisse_ tooma	107	T	0,198
laiali_ saatma	40	F	0,136	ilma_ olema	379	F	0,113	järele_ jääma	33	T	0,197
kõrvale_ jätma	62	T	0,135	järele_ mõtlema	213	T	0,112	esile_ kutsuma	316	T	0,195

Tabel 14. **50 kõrgeima ΔP(adverb|verb) väärtusega ühendit aja-, ilu- ja teaduskirjandustekstides.** Lühendid: sag – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: T = on õige ühendverb, F = ei ole õige ühendverb; ΔP(1|2) – ΔP(adverb|verb) väärtus.

ajakirjandus				**ilukirjandus**				**teaduskirjandus**			
adverb_ verb	**sag**	**yv**	**ΔP(1|2)**	**adverb_ verb**	**sag**	**yv**	**ΔP(1|2)**	**adverb_ verb**	**sag**	**yv**	**ΔP(1|2)**
sekka_ hällitama	1	F	1	ülal_ ekshibeerima	1	F	1	valla_ pääsima	1	F	1
tagant_ suskima	1	F	1	tagant_ punima	1	F	1	järel_ vantsima	1	F	1
püsti_ krapsama	1	F	1	tagant_ utjama	1	F	1	üleval_ semantiseeruma	1	F	1
laiali_ hoorama	1	F	1	valla_ loppuma	1	F	1	vahele_ tükkima	1	F	1

vahele_ pläristama	1	F	1	lahku_ translitereerima	1	F	1	taga_ lähnuma	1	F	1
kaasas_ tilpnema	1	F	0,999	peal_ ruiguma	1	F	1	ühte_ kollabseeruma	1	F	1
mööda_ sudima	1	F	0,999	kallale_ hassetama	1	F	1	püsti_ kohisema	1	F	1
järele_ reformeerima	1	F	0,999	alt_ kolmandama	1	F	1	mööda_ visioneerima	1	F	1
täis_ maruma	1	F	0,999	alt_ muduma	1	F	1	järele_ ahvima	1	F	1
täis_ roojama	1	F	0,999	üleval_ poositama	1	F	1	järele_ võsastuma	1	F	1
ilma_ etiketistama	1	F	0,999	pealt_ ilgema	1	F	1	järele_ önnistuma	1	F	1
ilma_ kekutama	1	F	0,999	pealt_ satuma	1	F	1	ringi_ meikima	1	F	1
ringi_ floresiensima	1	F	0,999	kaasas_ taastunnustama	1	F	0,999	ringi_ pöruma	1	F	1
ringi_ kooberdama	1	F	0,999	laiali_ sõõrduma	5	F	0,999	ringi_ sobrama	1	F	1
ringi_ melutsema	1	F	0,999	laiali_ kluksuma	1	F	0,999	ringi_ vurama	1	F	1
ringi_ roobeldama	1	F	0,999	laiali_ siia-sinama	1	F	0,999	lahti_ rulluma	3	T	1
ümber_ grupeeruma	1	F	0,999	eemale_ lootsima	1	F	0,999	lahti_ muukima	2	F	1
koos_ armeerima	1	F	0,999	valmis_ schoolima	1	F	0,999	lahti_ kasukkama	1	F	1
koos_ kudrutama	1	F	0,999	ilma_ jõutama	1	F	0,999	lahti_ korkima	1	T	1
koos_ kypsetama	1	F	0,999	mööda_ ehituma	1	F	0,999	kinni_ tukastama	2	F	1
valmis_ jäigendama	1	F	0,999	mööda_ neli-viima	1	F	0,999	kinni_ valendama	1	F	1
valmis_ plökerdama	1	F	0,999	mööda_ vuhkima	1	F	0,999	kõrvale_ põiklema	1	F	0,999
valmis_ sheigima	1	F	0,999	ümber_ plartseldama	2	F	0,999	peale_ illotuma	1	F	0,999
peale_ tarretuma	2	F	0,999	ümber_ defineeruma	1	F	0,999	ilma_ indemnitama	1	F	0,999
peale_ härduma	1	F	0,999	ümber_ parantsatama	1	F	0,999	ilma_ sandistama	1	F	0,999
lahti_ kirjutatama	1	F	0,998	ümber_ plumpsuma	1	F	0,999	maha_ kopuleeruma	1	F	0,999
lahti_ magneetima	1	F	0,998	ümber_ profileerima	1	F	0,999	maha_ kõmpima	1	T	0,999
lahti_ raspeldama	1	F	0,998	püsti_ ehituma	1	F	0,999	maha_ materdama	1	F	0,999
lahti_ suutduma	1	F	0,998	püsti_ kräuksatama	1	F	0,999	maha_ prassima	1	F	0,999
ligi_ taasinvesteerima	1	F	0,998	püsti_ looberdama	1	F	0,999	maha_ tolgotama	1	F	0,999
sisse_ logima	3	T	0,998	püsti_ virtsama	1	F	0,999	maha_ võtima	1	F	0,999

sisse_ canariensima	1	F	0,998	täis_ köettuma	1	F	0,998	ümber_ kiskuma	2	F	0,999
sisse_ desifintseerima	1	F	0,998	täis_ litsitama	1	F	0,998	ümber_ portima	1	F	0,999
sisse_ freesima	1	F	0,998	täis_ plätsutama	1	F	0,998	ümber_ rahvastuma	1	F	0,999
sisse_ lõõskama	1	F	0,998	otsa_ krüptima	1	F	0,998	ümber_ suupärastama	1	F	0,999
sisse_ melutsema	1	F	0,998	otsa_ tõsima	1	F	0,998	ümber_ taasäratama	1	F	0,999
sisse_ susserdama	1	F	0,998	järele_ kösima	1	F	0,998	sisse_ kasseerima	2	F	0,999
sisse_ talluma	1	F	0,998	järele_ lausima	1	F	0,998	sisse_ aerutama	1	F	0,999
sisse_ võmmutama	1	F	0,998	järele_ lõõgastama	1	F	0,998	sisse_ annoteerima	1	F	0,999
alla_ aetama	1	F	0,998	järele_ tšekkama	1	F	0,998	sisse_ kreemitama	1	F	0,999
alla_ laanima	1	F	0,998	ringi_ ehituma	1	F	0,998	sisse_ meldima	1	F	0,999
kinni_ graundima	1	F	0,998	ringi_ helklema	1	F	0,998	alla_ kriipsutama	15	T	0,999
kinni_ klammerdama	1	F	0,998	ringi_ lidisema	1	F	0,998	alla_ kantima	1	F	0,999
kinni_ mürtsatama	1	F	0,998	ringi_ lidristama	1	F	0,998	alla_ sähvatama	1	F	0,999
üles_ lebasklema	1	F	0,997	ringi_ lööberdama	1	F	0,998	alla_ väärtustatama	1	F	0,999
üles_ manageerima	1	F	0,997	ringi_ põrklema	1	F	0,998	üles_ miesima	1	F	0,999
üles_ nõmetsema	1	F	0,997	ringi_ tõmama	1	F	0,998	üles_ taaskiirguma	1	F	0,999
vastu_ margistama	1	F	0,997	peale_ hoopistykkima	1	F	0,998	üles_ upitama	1	F	0,999
vastu_ roosatama	1	F	0,997	peale_ komistuma	1	F	0,998	tagasi_ repatrieeruma	1	F	0,999
vastu_ rudima	1	F	0,997	peale_ pitserdama	1	F	0,998	tagasi_ saalima	1	F	0,999

Ajalehetekstide 50-st kõrgeima ΔP(verb|adverb) väärtusega ühendist on 14 sellised, mis EKSS-i loendis ei esine. Kuna ΔP(verb|adverb) väärtus kajastab, kuidas verbi esinemine sõltub adverbi esinemisest, siis kõrge ΔP(verb|adverb) väärtuse on saanud sellised õiged ühendverbid nagu *pärale jõudma*, *vahele jääma*, *lahku minema*. Nendes ühendites on adverbi ja verbi vahel tugev seos, mis näitab, et kui osalauses esineb ühendite adverbiline komponent, siis suure tõenäosusega esineb samas osalauses ka ühendi verbiline komponent. Näiteks kui osalauses esineb adverb *pärale*, siis on suur tõenäosus, et seal esineb verb *jõudma*. Ilukirjandustekstidest tuvastas ΔP(verb|adverb) 31

õiget ühendverbi (nt *pihta saama, vahele jääma*) ja teadustekstidest 28 õiget ühendverbi (nt *vastu võtma, läbi viima*). Järelikult tuvastab ΔP(verb|adverb) kõige paremini ajakirjandustekstidest õigeid ühendverbe, kõige halvemini aga teaduskirjandusest.

Kui vaadelda tulemusi, mis kirjeldavad, kuidas adverbi esinemine sõltub verbi esinemisest samas osalauses (ΔP(adverb|verb)), siis 50 kõrgeima väärtuse saanud ühendi hulgas on õigeid ühendverbe vähe. Põhjuseks on see, et kui ühend esineb vaid ühe korra ja selle verbiline komponent ei ole sage, siis on ühendi komponentide vahel väga tugev seos, mis tekib ka sellistel juhtudel, kui harvaesineva verbiga juhtub samas osalauses olema (juhuslik) adverb. Näiteks kui korpuses esineb vaid korra verb *suskima* ja samas osalauses on adverb *tagant*, siis on nende kahe sõna vahel tugev seos, mis kajastub selles, et kui esineb verb *suskima*, siis samas osalauses esineb alati ka adverb *tagant*.

ΔP(adverb|verb) tuvastab ajakirjandustekstidest vaid ühe õige ühendverbi (*sisse logima*) ja teaduskirjandusest neli õiget ühendverbi (*lahti rulluma, lahti korkima, maha kõmpima, alla kriipsutama*). Ilukirjandusest ei tuvasta ΔP(adverb|verb) mitte ühtegi õiget ühendverbi. Seega on ΔP(adverb|verb) kõige tulemuslikum just teaduskirjandusest õigete ühendverbide tuvastamisel.

Kokkuvõtlikult oleneb ΔP tulemus aga sellest, missugust seost vaadelda: ΔP(verb|adverb) tuvastab kõige paremini ühendverbe ajakirjandustekstidest, ΔP(adverb|verb) aga teadustekstidest.

3.1.3.3. Kokkuvõte asümmeetriliste mõõdikute tulemustest Tasakaalus korpuse põhjal

Kui võrrelda tinglikku tõenäosuse ja ΔP tulemusi, siis mõlemal juhul on tulemused paremad, kui vaadelda verbi esinemise sõltuvust adverbi esinemisest. Teisisõnu sellise seose vaatlemine on õigete ühendverbide tuvastamisel tulemuslikum. Vastupidise seose (adverbi sõltumine verbist) tugevuse mõõtmine aitab paremini tuvastada harva esinevaid ühendeid, mille verbiline komponent on samuti harv.

Tabel 15 esitab asümmeetriliste statistikute tuvastatud õigete ühendverbide arvud Tasakaalus korpuse tekstiklasside põhjal, kui vaadeldavate kandidaatpaaride arv 50.

Tabel 15. **Asümmeetriliste statistikute tuvastatud õigete ühendverbide arv aja-, ilu- ja teaduskirjandustekstides.**

tekstiliik	CP(verb|adverb)	CP(adverb|verb)	ΔP(verb|adverb)	ΔP(adverb|verb)
ajakirjandus	25	1	36	1
ilukirjandus	18	0	31	0
teaduskirjandus	18	5	28	4

50 kõrgeima statistiku väärtuse saanud kandidaatpaari seast kõikidest tekstiklassidest tuvastab ΔP(verb|adverb) rohkem õigeid ühendverbe kui CP (verb|adverb): ajakirjanduse korpusest tuvastab ΔP 36 ja tinglik tõenäosus 25 õiget ühendverbi, ilukirjandusest vastavalt 31 ja 18 ning teaduskirjandusest 28 ja 18 õiget ühendverbi. CP(adverb|verb) ja ΔP(adverb|verb) tuvastavad ajakirjandustekstidest ühe õige ühendverbi ja teaduskirjandusest vastavalt viis ja neli EKSS-i loendisse kuuluvat ühendit. Ilukirjandustekstidest ei tuvasta kumbki ühtegi õiget ühendverbi. Nii ei ole CP(adverb|verb) ja ΔP(adverb|verb) tulemuslikud õigete ühendverbide tuvastamisel, kuid nende tulemustes sisaldub informatsioon ühendi komponentide seoste kohta ja nad sobivad hästi harvaesinevate ühendite tuvastamiseks.

Seega ΔP(verb|adverb) tuvastab rohkem õigeid ühendverbe kui CP(verb|adverb). CP(adverb|verb) ja ΔP(adverb|verb) tulemusi võib aga pidada võrdseteks.

3.1.4. Sümmeetriliste ja asümmeetriliste mõõdikute tulemuste võrdlus Tasakaalus korpuse põhjal

Selleks et võrrelda, kas ja kuidas sümmeetriliste ja asümmeetriliste statistikute tulemused erinevad erinevatest tekstiklassidest õigete ühendverbide tuvastamisel, ning kas asümmeetriliste statistikute kasutamine on põhjendatud, võrdlen tuvastatud õigete ühendverbide loendeid. Aluseks võtan eespool esitatud tabelid, kus on iga mõõdiku 50 kõrgeima väärtusega ühendit, millest eraldan õiged ühendverbid ja koostan eraldi loetelud sümmeetriliste ja asümmeetriliste statistikute tulemustest. Loendites esineb iga õige ühendverb korra. Nende loetelude võrdlemisel ja vaatlemisel selgub, kas ja kui palju on selliseid õigeid ühendverbe, mida asümmeetrilised tuvastavad, kuid sümmeetrilised mitte.

Tabel 16 esitab tekstiklasside lõikes sümmeetriliste ja asümmeetriliste mõõdikute tuvastatud õigete ühendverbide arvud. Samuti nende ühendite arvud, mida

sümmeetrilised statistikud tuvastavad, kuid asümmeetrilised ei tuvasta, ning nende õigete ühendverbide arvud, mida asümmeetrilised tuvastavad, kuid sümmeetrilised ei tuvasta.

Tabel 16. **Sümmeetriliste ja asümmeetriliste mõõdikute tuvastatud õigete ühendverbide arv aja-, ilu- ja teaduskirjandustekstides.**

tekstiliik korpuses	**sümmeetrilised**	**asümmeetrilised**	**ainult sümmeetrilised**	**ainult asümmeetrilised**
ajakirjandus	83	37	69	23
ilukirjandus	79	31	63	15
teaduskirjandus	68	33	52	17

Ajakirjandustekstidest tuvastavad sümmeetrilised meetodid koos lihtsa sagedusloendiga kokku 83 erinevat õiget ühendverbi, asümmeetrilised statistikud aga 37 erinevat ühendverbi, millest 23 on sellised, mida sümmeetrilised mõõdikud ei tuvasta. Ilukirjandustekstidest tuvastavad sümmeetrilised meetodid koos lihtsa sagedusloendiga kokku 79 erinevat õiget ühendverbi, asümmeetrilised aga 31 erinevat ühendverbi, millest 15 on sellised, mida sümmeetrilised ei tuvasta. Teadustekstidest tuvastavad sümmeetrilised meetodid koos lihtsa sagedusloendiga kokku 68 erinevat õiget ühendverbi, asümmeetrilised statistikud aga 33, millest 17 on sellised, mida sümmeetrilised ei tuvasta.

Kuigi sümmeetrilised statistikud tuvastavad suure hulga selliseid ühendeid, mida asümmeetrilised ei tuvasta, tasub siiski asümmeetrilisi mõõdikuid ühendverbide tuvastamisel kasutada, sest kaks asümmeetrilist mõõdikut suudavad tuvastada kõikidest tekstiklassidest selliseid ühendverbe, mida viis sümmeetrilist ja lihtne sagedusloend ei tuvasta.

Siiski on selge, et kui välja valida vaid üks ja kõige edukam statistik ülesande lahendamiseks, siis selles töös kasutatud materjali põhjal on selleks sümmeetriline statistik t-skoor, mille 50 kõrgeima väärtusega ühendi seas on kõige rohkem õigeid ühendverbe (vt ptk 3.1.2.1).

Gries (2012) toob välja rea põhjuseid, miks korpuslingvistika peaks koostööd tegema kognitiivse keeleteaduse ja psühholingvistikaga. Ka siinne katse näitas, et parima tulemuse saamiseks on mõistlik kasutada sümmeetrilisi ja psühholoogiliselt tugevama alusega asümmeetrilisi statistikuid koos või kombineeritult, sest ainuüksi sümmeetrilisi

mõõdikuid kasutades jäävad tuvastamata asümmeetrilised seosed sõnade vahel. Järelikult võiks eesti keele ühendverbide tuvastamisel mõistlik olla parima sümmeetrilise statistiku t-skoori kombineerimine parima asümmeetrilise statistiku ΔP-ga. Niisamuti võiks sellesse kombinatsiooni lisada edukalt töötanud log-tõepära funktsiooni, lihtsa sagedusloendi aga ka tingliku tõenäosuse. Kuna korpuses sisalduv tekstiliik mõjutab mõnevõrra mõõdikute tulemusi, siis saab statistikute kombineerimisel jälgida iga tekstiliigi täpset statistikute paremusjärjestust, et iga tekstiliigi jaoks parim võimalik kombinatsioon välja töötada.

Statistikute kombineerimine, nende kombinatsioonide rakendamine ja tulemuste võrdlemine selle töö piiridesse ei mahu, kuid sellised katsed on tulevikus kindlasti vajalikud.

3.2. Tulemused sõltuvalt korpuse kasvust

See peatükk kirjeldab statistikute tulemusi ühendverbide tuvastamisel erineva suurusega korpustest. Eesmärk on välja selgitada, kas ja kuidas korpuse mahu suurenemine mõjutab mõõdikute tööd.

3.2.1. Lihtsa sagedusloendi täpsus ja saagis erineva suurusega korpuste lõikes

Selleks et vaadelda, kas ja kuidas valimi suurus mõjutab mõõdikute tulemusi, kasutan koondkorpuse ajakirjandustekste, millest siinsesse töösse kaasan 170 miljonit sõna. See sõnade hulk on jaotatud viieks osaks: esimene valim koosneb 5 miljonist sõnast ja sisaldab ainult Tasakaalus korpusesse kuuluvaid ajakirjandustekste, järgmised osad koosnevad vastavalt 10, 20, 70 ja 170 miljonist sõnast. Lisaks statistikute tulemustele vaatlen ka lihtsa sagedusloendi tulemusi.

Tabel 17 esitab ülevaate lihtsa sagedusloendi tulemustest erineva suurusega korpustest ühendverbide tuvastamisel.

Tabel 17. **Sageduse täpsus ja saagis erineva suurusega ajakirjanduskorpustes.**

sõnu	osalauseid	kandidaatpaare	õigeid ühendverbe	täpsus	saagis
5 000 000	707 979	13 141	1351	10,3%	77,8%
10 000 000	1 410 474	18 545	1459	7,9%	84,0%
20 000 000	2 823 255	26 268	1532	5,8%	88,2%
70 000 000	9 640 426	46 863	1628	3,5%	93,7%
170 000 000	24 322 394	67 558	1676	2,5%	96,5%

Selgub, et näiteks 20 000 000 sõnast koosneva ajakirjandustekstide valimi alusel genereeritud kandidaatpaaride arv on 26 268 ning tuvastatud õigete ühendverbide arv on 1532, mis on 88,2% kõikidest õigetest ühendverbidest. Kõikidest genereeritud kandidaatpaaridest on sellisel juhul 5,8% EKSS-i kuuluvad ühendverbid. Ajakirjandustekstidest genereeriti kokku 67 558 kandidaatpaari, millest 1676 kuuluvad EKSS-i loendisse. See tähendab, et tuvastati 96,5% kõigist EKSS-i kuuluvatest ühenditest ehk tuvastamata jäi 61 õiget ühendverbi. Suure hulga kandidaatpaaride tõttu on sageduse täpsus aga madal (2,5%).

3.2.2. Sümmeetriliste mõõdikute ja lihtsa sagedusloendi tulemused erineva suurusega korpuste põhjal

See peatükk esitab ülevaate sümmeetriliste mõõdikute ja lihtsa sagedusloendi tulemustest ühendverbide tuvastamisel erineva suurusega korpustest. Eesmärk on vaadelda, kas ja kuidas muutuvad statistikute tulemused korpuse suurenemisega.

3.2.2.1. t-skoori tulemused erineva suurusega korpuste põhjal

Tabel 18 esitab kahanevalt 50 kõrgeima t-skoori väärtusega sõnapaari 170 miljoni sõna suuruses ajakirjandustekstide korpuses.

Tabel 18. **50 kõrgeima t-skoori väärtusega ühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; t-skoor – t-skoori väärtus.

adverb	f1	verb	f2	sagedus	yv	t-skoor
vastu	76337	võtma	322185	28435	TRUE	162,630
ette	114045	nägema	200795	26025	TRUE	155,486
välja	297714	tulema	567778	28117	TRUE	126,235
kaasa	75124	tooma	142990	16097	TRUE	123,393
kinni	60871	pidama	702611	16411	TRUE	114,379
välja	297714	kuulutama	32307	13196	TRUE	111,431
kokku	150217	leppima	26637	12310	TRUE	109,468
läbi	91173	viima	116301	12750	TRUE	109,055
alla	59223	kirjutama	132506	12218	TRUE	107,616
välja	297714	andma	335131	18290	TRUE	104,909
ette	114045	kujutama	25135	11132	TRUE	104,391
ette	114045	võtma	322185	13163	TRUE	101,563

ette	114045	valmistama	34115	9834	TRUE	97,553
tagasi	100784	tulema	567778	13053	TRUE	93,657
alla	59223	jääma	360888	10308	TRUE	92,873
üle	134583	andma	335131	11941	TRUE	92,305
kaasa	75124	aitama	74407	8905	TRUE	91,931
maha	74140	müüma	73763	8435	TRUE	89,394
välja	297714	töötama	87160	9174	TRUE	84,642
kaasa	75124	võtma	322185	8969	FALSE	84,197
ära	176797	kasutama	132094	8886	TRUE	84,080
maha	74140	võtma	322185	8907	TRUE	83,971
kinni	60871	maksma	153513	7735	TRUE	83,580
ilma	21833	jääma	360888	7505	TRUE	82,892
välja	297714	tooma	142990	10029	TRUE	82,668
välja	297714	nägema	200795	11105	TRUE	82,057
vastu	76337	pidama	702611	10674	TRUE	81,971
ära	176797	võtma	322185	10527	TRUE	79,776
edasi	75469	lükkama	28633	6382	TRUE	78,775
välja	297714	pakkuma	102085	8370	TRUE	77,830
vahele	10777	jääma	360888	6249	TRUE	77,028
välja	297714	vahetama	27883	6543	TRUE	76,669
ära	176797	tegema	598768	12690	TRUE	74,014
üles	64480	kutsuma	63938	5474	TRUE	71,695
tagasi	100784	lükkama	28633	5364	TRUE	71,619
tagasi	100784	astuma	52604	5531	TRUE	71,440
välja	297714	valima	86366	7011	TRUE	71,106
maha	74140	jääma	360888	7033	TRUE	70,746
kokku	150217	puutuma	12256	5135	TRUE	70,603
välja	297714	selgitama	71875	6497	TRUE	69,689
üle	134583	jääma	360888	8218	TRUE	68,625
ette	114045	heitma	17617	4806	TRUE	68,134
kaasa	75124	tegema	598768	7858	TRUE	67,782
üle	134583	minema	365466	8105	TRUE	67,565
edasi	75469	minema	365466	6620	TRUE	67,426
üle	134583	vaatama	114220	5690	TRUE	67,054
läbi	91173	käima	172564	5669	TRUE	66,702
üles	64480	astuma	52604	4630	TRUE	65,995
kokku	150217	hoidma	76157	5247	TRUE	65,943
ära	176797	tundma	129460	6059	TRUE	65,750

t-skoori 50-st kõrgeima väärtusega sõnapaarist on vaid üks selline (*kaasa võtma*), mida EKSS-i nimistus ei ole. Kui võrrelda seda tulemust Tasakaalus korpuse ajakirjandustekstidest saadud esimese 50 kõrgeima t-skoori väärtuse saanud sõnaühendiga (vt tabel 4), siis on see arv ühe võrra vähenenud, mistõttu võib oletada, et korpuse suuruse kasvamine parandab mõnevõrra t-skoori tulemusi ühendverbide tuvastamisel ajakirjandustekstidest.

3.2.2.2. MI tulemused erineva suurusega korpuste põhjal

Tabel 19 esitab kahanevalt 50 kõrgeima MI väärtusega sõnapaari 170 miljoni sõna suuruses ajakirjandustekstide korpuses.

Tabel 19. **50 kõrgeima MI väärtusega ühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses; f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; MI – MI väärtus.

adverb	f1	verb	f2	sagedus	yv	MI
sekka	1124	jõratama	1	1	FALSE	14,401
sekka	1124	mahakooruma	1	1	FALSE	14,401
sekka	1124	taas-taas-taasavaldama	1	1	FALSE	14,401
ühte	2132	kanseldama	1	1	FALSE	13,478
järel	2166	mitte-kunagi-armastama	1	1	FALSE	13,455
tagant	2506	turgima	1	1	FALSE	13,245
kallale	3189	õppi-matma	1	1	FALSE	12,897
kallale	3189	tromama	1	1	FALSE	12,897
valla	3269	portesteerima	1	1	FALSE	12,861
sekka	1124	needistama	3	1	FALSE	12,816
sekka	1124	tšekkama	7	2	FALSE	12,594
pärale	841	muganema	5	1	FALSE	12,498
alt	4217	delegeeruma	1	1	FALSE	12,494
alt	4217	tsirukleerima	1	1	FALSE	12,494
üleval	5143	walitsema	1	1	FALSE	12,207
taha	1095	praeguma	5	1	FALSE	12,117
tasa	2188	posisema	3	1	FALSE	11,855
taga	7767	garaaxima	1	1	FALSE	11,613
taga	7767	pusletama	1	1	FALSE	11,613
taga	7767	nõduma	1	1	FALSE	11,613
alt	4217	näpsima	2	1	FALSE	11,494
eemale	8654	seminaritsema	1	1	FALSE	11,457

eemale	8654	otsas/muigama	1	1	FALSE	11,457
eemale	8654	jää-minema	1	1	FALSE	11,457
eemale	8654	hoidsima	1	1	FALSE	11,457
eemale	8654	põnklema	1	1	FALSE	11,457
eemale	8654	neljutama	1	1	FALSE	11,457
eemale	8654	timpima	1	1	FALSE	11,457
külge	2997	pookima	416	136	TRUE	11,374
püsti	9341	ballastima	1	1	FALSE	11,346
laiali	9712	kon-marineerima	1	1	FALSE	11,290
laiali	9712	toosistama	1	1	FALSE	11,290
laiali	9712	lammutatama	1	1	FALSE	11,290
laiali	9712	peks-matma	1	1	FALSE	11,290
sekka	1124	kekkama	9	1	FALSE	11,231
sekka	1124	hällitama	9	1	FALSE	11,231
üleval	5143	ohatama	2	1	FALSE	11,207
ühte	2132	pattama	10	2	FALSE	11,156
pealt	5394	leitima	2	1	FALSE	11,139
tagant	2506	utsitama	400	86	TRUE	11,027
külge	2997	aplitseerima	4	1	FALSE	10,986
kallale	3189	õssitama	4	1	FALSE	10,897
otsa	13292	pilguma	1	1	FALSE	10,838
otsa	13292	lõppetama	1	1	FALSE	10,838
otsa	13292	omima	1	1	FALSE	10,838
mööda	13752	pargi/lõikama	1	1	FALSE	10,788
mööda	13752	melodeklameerima	1	1	FALSE	10,788
mööda	13752	hargi/roitma	1	1	FALSE	10,788
mööda	13752	rümama	1	1	FALSE	10,788
sekka	1124	taaskehastuma	13	1	FALSE	10,701

MI tulemused on ootuspärased, kui arvestada väidet, et MI on loodud madala esinemissagedustega ühendite leidmiseks ja tõstab esile harvaesinevaid ühendeid, mille komponendid esinevad samuti harva. 50 kõrgeima MI väärtuse saanud sõnapaaride seas on enamus sellised, mis esinevad korpuses vaid ühe korra (nt *sekka needistama*). Nende hulgas on vaid kaks EKSS-i nimistusse kuuluvat ühendverbi (*külge pookima*, *tagant utsitama*), mis esinevad korpuses suhteliselt harva. Võrreldes Tasakaalus korpuse ajakirjandustekstidest saadud 50 kõrgema MI väärtuse saanud ühendiga (vt tabel 5) on õigete ühendverbide arv vähenenud nelja võrra. Järelikult korpuse suuruse kasvamine halvendab MI tulemusi ühendverbide tuvastamisel ajakirjandustekstidest.

3.2.2.3. Hii-ruut-statistiku tulemused erineva suurusega korpuste põhjal

Tabel 20 esitab kahanevalt 50 kõrgeima hii-ruut-statistiku väärtusega sõnapaari 170 miljoni sõna suuruses ajakirjandustekstide korpuses (väärtused on läbi korrutatud (–1)-ga, juhul kui O < E).

Tabel 20. **50 kõrgeima hii-ruut-statistiku väärtusega ühendit ajakirjandustekstides.**
Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses, sagedus – ühendi koosesinemissagedus samas osalauses, yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; hii-ruut – hii-ruut-statistiku väärtus.

adverb	f1	verb	f2	sagedus	yv	hii-ruut
kallale	3189	tungima	8817	1385	FALSE	1657370,556
eemale	8654	peletama	2535	1083	FALSE	1298806,643
ette	114045	kujutama	25135	11132	TRUE	1035244,486
kokku	150217	leppima	26637	12310	TRUE	903229,861
vastu	76337	võtma	322185	28435	TRUE	756097,657
ette	114045	nägema	200795	26025	TRUE	677009,296
ette	114045	valmistama	34115	9834	TRUE	588641,729
kaasa	75124	tooma	142990	16097	TRUE	559953,187
esile	15517	tõstma	46220	3920	TRUE	514618,721
alt	4217	vedama	24469	1455	TRUE	496693,677
edasi	75469	lükkama	28633	6382	TRUE	447681,401
alla	59223	kirjutama	132506	12218	TRUE	442045,219
edasi	75469	lükkuma	2945	2004	FALSE	436898,357
välja	297714	kuulutama	32307	13196	TRUE	420043,019
kokku	150217	põrkama	5227	3595	TRUE	395714,021
ümber	26581	lükkama	28633	3402	TRUE	363912,430
külge	2997	pookima	416	136	TRUE	360609,836
läbi	91173	viima	116301	12750	TRUE	350808,913
kokku	150217	puutuma	12256	5135	TRUE	340431,568
kaasa	75124	aitama	74407	8905	TRUE	329492,093
maha	74140	müüma	73763	8435	TRUE	301621,804
kõrvale	15717	hiilima	2250	639	TRUE	279768,098
ette	114045	heitma	17617	4806	TRUE	271558,017
taga	7767	ajama	46586	2019	TRUE	270593,813
taga	7767	kiusama	1640	352	TRUE	235976,755
vahele	10777	jääma	360888	6249	TRUE	235464,844
tagasi	100784	lükkama	28633	5364	TRUE	233137,644
esile	15517	kutsuma	63938	3074	TRUE	226289,607
lahti	34458	harutama	613	434	TRUE	216332,652

tagant	2506	utsitama	400	86	TRUE	179307,192
kinni	60871	nabima	616	519	TRUE	174126,646
otsa	13292	sõitma	115057	3339	TRUE	171601,011
esile	15517	kerkima	19624	1462	FALSE	168059,645
üles	64480	kutsuma	63938	5474	TRUE	166880,913
ilma	21833	jääma	360888	7505	TRUE	161725,482
kõrvale	15717	põiklema	304	175	FALSE	155650,231
vastu	76337	vaidlema	8140	1961	TRUE	147137,137
esile	15517	tooma	142990	3732	TRUE	146257,719
üles	64480	astuma	52604	4630	TRUE	145295,975
laiali	9712	laotama	1001	241	FALSE	144892,936
lahti	34458	muukima	462	308	FALSE	144527,964
kokku	150217	varisema	2771	1580	TRUE	143630,077
kinni	60871	maksma	153513	7735	TRUE	141892,297
alla	59223	kukkuma	32579	3361	FALSE	136272,934
välja	297714	lülitama	6938	3463	TRUE	136076,888
alla	59223	neelama	3963	1154	TRUE	136062,812
peale	33718	suruma	13933	1636	TRUE	135581,957
tagasi	100784	astuma	52604	5531	TRUE	130323,732
kinni	60871	pidama	702611	16411	TRUE	126045,884
üles	64480	ehitama	52283	4239	TRUE	121887,617

50 kõrgeima hii-ruut-statistiku väärtuse saanud sõnapaari hulgas on nii kõrge (nt *vastu võtma*, *ette nägema*) kui ka suhteliselt madala (nt *tagant utsitama*) sagedusega ühendverbe. Kuigi hii-ruut-statistik eelistab madala esinemissagedusega sõnu ja ühendeid, saavad kõrged väärtused ka paljud õiged ühendverbid. Tabelis 20 esitatud ühendist on kaheksa sellised, mis EKSS-i loendisse ei kuulu (nt *kallale tungima*, *eemale peletama*). 5 miljoni sõna suuruse korpuse põhjal saadud 50 kõrgeima hii-ruut-statistiku väärtusega sõnapaari seas on neid üheksa (vt tabel 6). Korpuse suurenemine parandab hii-ruut-statistiku tulemusi ning ka 170 miljoni sõna suurusest korpusest tuvastab hii-ruut-statistik ühendverbe halvemini kui t-skoor, kuid paremini kui MI.

3.2.2.4. Log-tõepära funktsiooni tulemused erineva suurusega korpuste põhjal

Tabel 21 esitab kahanevalt 50 kõrgeima log-tõepära funktsiooni väärtusega sõnapaari 170 miljoni sõna suuruses ajakirjandustekstide korpuses (väärtused on läbi korrutatud (–1)-ga, juhul kui O < E).

Tabel 21. **50 kõrgeima log-tõepära funktsiooni väärtusega ühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; log-tõepära – log-tõepära funktsiooni väärtus.

adverb	**f1**	**verb**	**f2**	**sagedus**	**yv**	**log-tõepära**
vastu	76337	võtma	322185	28435	TRUE	148821,694
ette	114045	nägema	200795	26025	TRUE	131946,207
kaasa	75124	tooma	142990	16097	TRUE	89789,560
kokku	150217	leppima	26637	12310	TRUE	89664,217
ette	114045	kujutama	25135	11132	TRUE	86113,704
välja	297714	kuulutama	32307	13196	TRUE	73543,327
alla	59223	kirjutama	132506	12218	TRUE	68705,446
ette	114045	valmistama	34115	9834	TRUE	65570,984
läbi	91173	viima	116301	12750	TRUE	64569,727
kaasa	75124	aitama	74407	8905	TRUE	49887,973
kinni	60871	pidama	702611	16411	TRUE	48297,773
maha	74140	müüma	73763	8435	TRUE	46632,172
edasi	75469	lükkama	28633	6382	TRUE	44016,218
välja	297714	tulema	567778	28117	TRUE	38669,778
vahele	10777	jääma	360888	6249	TRUE	38200,100
kokku	150217	puutuma	12256	5135	TRUE	35838,768
ilma	21833	jääma	360888	7505	TRUE	35676,808
ette	114045	võtma	322185	13163	TRUE	35371,844
alla	59223	jääma	360888	10308	TRUE	33768,313
kinni	60871	maksma	153513	7735	TRUE	33037,772
esile	15517	tõstma	46220	3920	TRUE	31964,965
tagasi	100784	lükkama	28633	5364	TRUE	31706,701
ette	114045	heitma	17617	4806	TRUE	31216,671
kokku	150217	põrkama	5227	3595	TRUE	30191,417
üles	64480	kutsuma	63938	5474	TRUE	28339,401
välja	297714	vahetama	27883	6543	TRUE	27891,575
välja	297714	andma	335131	18290	TRUE	27628,295
ümber	26581	lükkama	28633	3402	TRUE	26022,501
tagasi	100784	astuma	52604	5531	TRUE	25990,677
üle	134583	andma	335131	11941	TRUE	25404,643
tagasi	100784	tulema	567778	13053	TRUE	24747,417
kaasa	75124	võtma	322185	8969	FALSE	24584,362
ära	176797	kasutama	132094	8886	TRUE	24547,437
maha	74140	võtma	322185	8907	TRUE	24519,483
välja	297714	töötama	87160	9174	TRUE	24279,096

üles	64480	astuma	52604	4630	TRUE	24170,255
kinni	60871	hoidma	76157	4672	TRUE	21540,367
esile	15517	tooma	142990	3732	TRUE	21450,557
üles	64480	ehitama	52283	4239	TRUE	21396,797
esile	15517	kutsuma	63938	3074	TRUE	21291,317
välja	297714	lülitama	6938	3463	TRUE	21001,781
otsa	13292	sõitma	115057	3339	TRUE	20957,072
kaasa	75124	lööma	53563	4185	TRUE	19532,149
edasi	75469	lükkuma	2945	2004	FALSE	19516,999
maha	74140	laskma	76128	4639	TRUE	19506,277
välja	297714	tooma	142990	10029	TRUE	19189,916
alla	59223	kukkuma	32579	3361	FALSE	19149,165
tagasi	100784	pöörduma	26413	3646	FALSE	19116,860
välja	297714	pakkuma	102085	8370	TRUE	18285,971
vastu	76337	pidama	702611	10674	TRUE	17840,647

50-st kõrgeima log-tõepära funktsiooni väärtusega ühendist neli ei ole EKSS-i loendis (*kaasa võtma*, *edasi lükkuma, alla kukkuma, tagasi pöörduma*). Tasakaalus korpuse 5 miljoni sõna suuruse ajakirjandustekstide osa põhjal leitud 50-st kõrgeima log-tõepära funktsiooni väärtuse saanud ühendist ei kuulu EKSS-i loendisse vaid kolm (vt tabel 7). Seega võib oletada, et korpuse mahu suurenemine halvendab log-tõepära funktsiooni tulemusi ajakirjandustekstidest ühendverbide tuvastamisel, kuid log-tõepära funktsioon tuvastab ka 170 miljoni sõna suurusest korpusest ühendverbe paremini kui hii-ruut-statistik ja MI.

3.2.2.5. MS-i tulemused erineva suurusega korpuste põhjal

Tabel 22 esitab kahanevalt 50 kõrgeima MS-i väärtusega sõnapaari 170 miljoni sõna suuruses ajakirjandustekstide korpuses.

Tabel 22. **50 kõrgeima MS-i väärtusega ühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; MS – MS-i väärtus.

adverb	**f1**	**verb**	**f2**	**sagedus**	**yv**	**MS**
kallale	3189	tungima	8817	1385	FALSE	0,157
ette	114045	nägema	200795	26025	TRUE	0,130
eemale	8654	peletama	2535	1083	FALSE	0,125

ümber	26581	lükkama	28633	3402	TRUE	0,119
kaasa	75124	aitama	74407	8905	TRUE	0,119
maha	74140	müüma	73763	8435	TRUE	0,114
kaasa	75124	tooma	142990	16097	TRUE	0,113
läbi	91173	viima	116301	12750	TRUE	0,110
ette	114045	kujutama	25135	11132	TRUE	0,098
alla	59223	kirjutama	132506	12218	TRUE	0,092
vastu	76337	võtma	322185	28435	TRUE	0,088
ette	114045	valmistama	34115	9834	TRUE	0,086
üles	64480	kutsuma	63938	5474	TRUE	0,085
esile	15517	tõstma	46220	3920	TRUE	0,085
edasi	75469	lükkama	28633	6382	TRUE	0,085
kokku	150217	leppima	26637	12310	TRUE	0,082
esile	15517	kerkima	19624	1462	FALSE	0,075
üles	64480	astuma	52604	4630	TRUE	0,072
üles	64480	ehitama	52283	4239	TRUE	0,066
kinni	60871	hoidma	76157	4672	TRUE	0,061
maha	74140	laskma	76128	4639	TRUE	0,061
alt	4217	vedama	24469	1455	TRUE	0,059
alla	59223	kukkuma	32579	3361	FALSE	0,057
kaasa	75124	lööma	53563	4185	TRUE	0,056
tagasi	100784	astuma	52604	5531	TRUE	0,055
välja	297714	andma	335131	18290	TRUE	0,055
tagasi	100784	lükkama	28633	5364	TRUE	0,053
kinni	60871	maksma	153513	7735	TRUE	0,050
tagant	2506	tõukama	4014	202	TRUE	0,050
ära	176797	kasutama	132094	8886	TRUE	0,050
välja	297714	tulema	567778	28117	TRUE	0,050
peale	33718	suruma	13933	1636	TRUE	0,049
esile	15517	kutsuma	63938	3074	TRUE	0,048
külge	2997	kleepima	2415	143	TRUE	0,048
eemale	8654	tõrjuma	8584	408	FALSE	0,047
ringi	23320	liikuma	39263	1803	FALSE	0,046
külge	2997	pookima	416	136	TRUE	0,045
taga	7767	kiusama	1640	352	TRUE	0,045
sisse	53046	seadma	32777	2391	TRUE	0,045
taga	7767	nutma	6249	350	TRUE	0,045
maha	74140	jätma	101181	4504	TRUE	0,045
välja	297714	kuulutama	32307	13196	TRUE	0,044
taga	7767	ajama	46586	2019	TRUE	0,043
üle	134583	vaatama	114220	5690	TRUE	0,042

ette	114045	heitma	17617	4806	TRUE	0,042
sisse	53046	astuma	52604	2168	TRUE	0,041
ette	114045	võtma	322185	13163	TRUE	0,041
kõrvale	15717	hiilima	2250	639	TRUE	0,041
tagant	2506	torkima	982	101	TRUE	0,040
üles	64480	seadma	32777	2556	TRUE	0,040

Ka 50 kõrgeima MS-i väärtuse saanud sõnapaaride seas on enamus EKSS-i loendisse kuuluvad ühendverbid. Sinna mitte kuuluvatest sõnapaaridest, mida on kokku kuus, on kõrge väärtuse saanud sellised paarid nagu *kallale tungima*, *eemale peletama*. 50-st kõrgeima MS-i väärtusega Tasakaalus korpuse ajakirjandustekstidest tuvastatud ühendist ei kuulu EKSS-i loendisse seitse sõnapaari (vt tabel 8). Järelikult parandab korpuse suurenemine MS-i tulemusi, kuid need on ka 170 miljoni sõna suuruse korpuse korral halvemad kui t-skoori ja log-tõepära funktsiooni tulemused.

3.2.2.6. Lihtsa sagedusloendi tulemused erineva suurusega korpuste põhjal

Tabel 23 esitab kahanevalt 50 sagedasemat sõnapaari 170 miljoni sõna suuruses ajakirjandustekstide korpuses.

Tabel 23. **50 sagedasemat sõnaühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb.

adverb	**f1**	**verb**	**f2**	**sagedus**	**yv**
välja	297714	olema	10739735	52328	FALSE
üle	134583	olema	10739735	34501	FALSE
kokku	150217	olema	10739735	33988	FALSE
ära	176797	olema	10739735	31531	FALSE
vastu	76337	võtma	322185	28435	TRUE
välja	297714	tulema	567778	28117	TRUE
ette	114045	nägema	200795	26025	TRUE
ette	114045	olema	10739735	25984	FALSE
läbi	91173	olema	10739735	23009	FALSE
valmis	30479	olema	10739735	21917	FALSE
välja	297714	andma	335131	18290	TRUE
kinni	60871	pidama	702611	16411	TRUE
kaasa	75124	tooma	142990	16097	TRUE

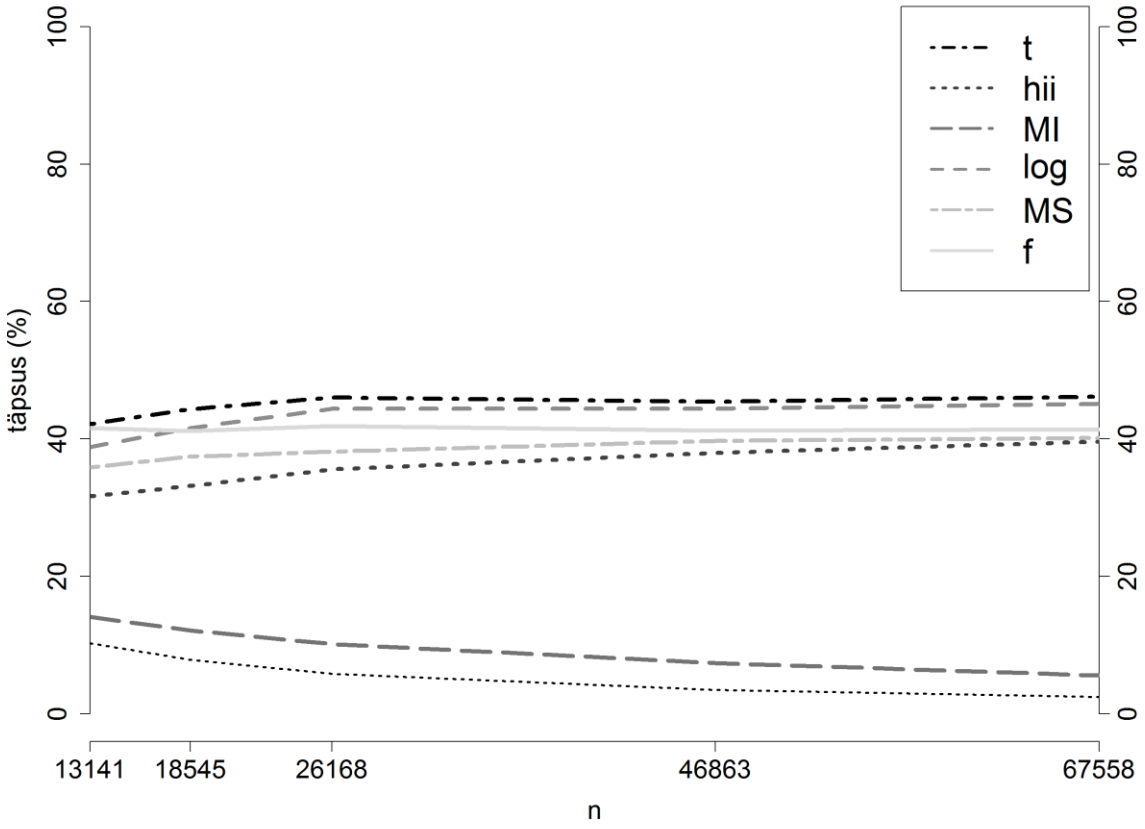
tagasi	100784	olema	10739735	15910	FALSE
kokku	150217	saama	1057514	13199	TRUE
välja	297714	kuulutama	32307	13196	TRUE
vastu	76337	olema	10739735	13167	FALSE
ette	114045	võtma	322185	13163	TRUE
tagasi	100784	tulema	567778	13053	TRUE
läbi	91173	viima	116301	12750	TRUE
ära	176797	tegema	598768	12690	TRUE
ligi	44960	olema	10739735	12346	FALSE
kokku	150217	leppima	26637	12310	TRUE
alla	59223	kirjutama	132506	12218	TRUE
üles	64480	olema	10739735	12151	FALSE
maha	74140	olema	10739735	12027	FALSE
üle	134583	andma	335131	11941	TRUE
ette	114045	kujutama	25135	11132	TRUE
välja	297714	nägema	200795	11105	TRUE
kinni	60871	olema	10739735	11047	FALSE
alla	59223	olema	10739735	10907	FALSE
sisse	53046	olema	10739735	10707	FALSE
vastu	76337	pidama	702611	10674	TRUE
ära	176797	võtma	322185	10527	TRUE
alla	59223	jääma	360888	10308	TRUE
kaasa	75124	olema	10739735	10151	FALSE
välja	297714	tooma	142990	10029	TRUE
edasi	75469	olema	10739735	9916	FALSE
välja	297714	saama	1057514	9847	TRUE
ette	114045	valmistama	34115	9834	TRUE
tagasi	100784	saama	1057514	9633	TRUE
koos	27259	olema	10739735	9566	FALSE
välja	297714	töötama	87160	9174	TRUE
kaasa	75124	võtma	322185	8969	FALSE
maha	74140	võtma	322185	8907	TRUE
kaasa	75124	aitama	74407	8905	TRUE
ära	176797	kasutama	132094	8886	TRUE
täis	16360	olema	10739735	8805	FALSE
välja	297714	pidama	702611	8760	TRUE
üle	134583	saama	1057514	8585	TRUE

Tabelist 23 selgub, et 50-st kõige sagedasemast ühendist on 20 mitte-ühendverbid. Tasakaalus korpuse ajakirjanduskorpuse 5 miljoni sõna suuruse valimi põhjal tuvastatud 50 sagedasema ühendi loendis on selliseid 21 (vt tabel 9). Võib öelda, et korpuse mahu kasvamine ei muuda oluliselt lihtsa sagedusloendi tulemusi.

3.2.2.7. Meetodite täpsused ja saagised erineva suurusega korpuste põhjal

Kui võrrelda 5 miljoni sõna ja 170 miljoni sõna suuruste ajakirjanduskorpuste põhjal statistikuid 50 kõige kõrgema statistiku väärtuse saanud kandidaatpaari alusel, siis võib oletada, et t-skoori, hii-ruut-statistiku, MS-i ja lihtsa sagedusloendi tulemused paranevad korpuse suuruse kasvamisega. Vastupidise tulemuse tekitab korpuse mahu suurenemine log-tõepära funktsiooni ja MI tulemustele. Lihtsa sagedusloendi tulemusi korpuse kasvamine ei näi oluliselt mõjutavat.

Seda, kas mõõdikute tulemused muutuvad korpuse suurenedes, kui arvesse võtta rohkem kandidaatpaare, illustreerib joonis 4, mis esitab sümmeetriliste mõõdikute ja lihtsa sagedusloendi täpsuse muutuse 2000 kandidaatpaari seas ühendverbide tuvastamisel erineva suurusega korpustest. Joonise x-telg märgib korpuse suuruse kasvamist erineva suurusega korpusest genereeritud kandidaatpaaride arvu näol, y-telg märgib mõõdikute täpsust, kui arvesse võtta esimesed 2000 nende kõrgeima väärtuse saanud kandidaatpaari. Peenike punktiirjoon joonisel tähistab algtaseme täpsust erineva suurusega korpustes.



Joonis 4. **Sümmeetriliste mõõdikute ja lihtsa sagedusloendi täpsus 2000 kandidaatpaari seas erineva suurusega ajakirjanduskorpustes.**

Joonis 4 näitab, et kõige enam mõjutab 2000 kandidaatpaari hulgas korpuse kasv hii-ruut-statistiku tulemust, sest selle täpsus tõuseb korpuse suurenemisega umbes 30%-lt 40%-ni. Samuti paranevad mõnevõrra t-skoori, log-tõepära funktsiooni ja MS-i tulemused. Lihtsa sagedusloendi tulemused korpuse mahu suurenemisega oluliselt ei muutu. MI tulemused halvenevad korpuse mahu suurendamisega. Eelnevalt tehtud oletused peavad üldiselt paika, vaid log-tõepära funktsiooni kohta tehtud oletus osutus valeks: log-tõepära funktsiooni tulemused siiski paranevad korpuse kasvuga.

Meetodite kõrvutamisel algtaseme täpsusega selgub, et kõikide mõõdikute täpsused on korpuse mahu suurenemisel kõrgemad kui algtaseme täpsus. Kui vaadelda esimest 2000 kõrgeima statistikute väärtustega kandidaatpaari, siis üldjoontes on kõik mõõdikud eesti keele ühendverbide tuvastamisel tulemuslikud. Samas paremusjärjestus mõnevõrra

muutub: kui 5 miljoni sõna suuruse korpuse korral on lihtne sagedusloend log-tõepära funktsioonist efektiivsem, siis suuremate korpuste korral on log-tõepära funktsiooni täpsus lihtsa sagedusloendi omast kõrgem. Ka võib öelda, et mida suurem on korpus, seda sarnasemad on lihtsa sagedusloendi, MS-i ja hii-ruut-statistiku tulemused. t-skoori täpsus on kõikide korpuste korral kõrgeim, kuid suurema korpuse puhul on log-tõepära tulemused t-skoori tulemustega sarnasemad kui väiksema korpuse korral.

Täpsema ülevaate korpuse mahu suurenemise mõjust mõõdikute tulemustele ja paremusjärjestusele saab tabelist 24, kus on esitatud sümmeetriliste statistikute ja lihtsa sagedusloendi täpsused ja saagised ühendverbide tuvastamisel erineva suurusega korpustest 100, 1000 ja 2000 kõrgeima väärtuse saanud sõnapaari seas.

Tabel 24. **Sümmeetriliste statistikute ja lihtsa sagedusloendi täpsused ja saagised erineva suurusega ajakirjanduskorpustes.**

Lühendid: stat – statistik, t – t-skoor, hii – hii-ruut-statistik, log – log-tõepära funktsioon, sag – lihtne sagedusloend.

stat	**meetod**	**5 000 000**			**20 000 000**			**170 000 000**		
		n=100	n=1000	n=2000	n=100	n=1000	n=2000	n=100	n=1000	n=2000
t	täpsus	95,0%	62,6%	42,2%	96,0%	64,5%	46,0%	95,0%	64,2%	46,2%
	saagis	5,5%	36,0%	48,6%	5,5%	37,1%	53,0%	5,5%	37,0%	53,1%
MI	täpsus	9,0%	11,9%	14,1%	8,0%	9,2%	10,2%	4,0%	4,1%	5,6%
	saagis	0,5%	6,9%	16,2%	0,5%	5,3%	11,7%	0,2%	2,4%	6,4%
hii	täpsus	71,0%	40,5%	31,6%	73,0%	48,4%	35,5%	79,0%	54,9%	39,6%
	saagis	4,1%	23,3%	36,4%	4,2%	27,9%	40,9%	4,5%	31,6%	45,6%
log	täpsus	88,0%	60,4%	38,8%	90,0%	63,0%	44,4%	88,0%	62,6%	45,1%
	saagis	5,1%	34,8%	44,7%	5,2%	36,3%	51,1%	5,1%	36,0%	52,0%
MS	täpsus	80,0%	51,0%	35,9%	86,0%	54,7%	38,1%	83,0%	56,2%	40,2%
	saagis	4,6%	29,4%	41,3%	5,0%	31,5%	43,9%	4,8%	32,4%	46,2%
sag	täpsus	73,0%	56,9%	41,6%	73,0%	59,0%	41,9%	73,0%	57,9%	41,4%
	saagis	4,2%	32,8%	47,9%	4,2%	34,0%	48,2%	4,2%	33,3%	47,7%

Tabelis 24 esitatud tulemused kinnitavad, et t-skooril on kõige paremad tulemused ühendverbide tuvastamisel erineva suurusega korpustest. Eelnevalt tehtud oletus, et t-skoori tulemused paranevad korpuse kasvades, kehtib, kui võtta võrdluse aluseks 2000 kõrgeima t-skoori väärtusega kandidaatpaari. Kui 5 miljoni sõna suuruse korpuse puhul on t-skoori täpsus 42,2%, siis 170 miljoni sõna suuruse korpuse korral on see 46,2%. Ka saagis paraneb 48,6%-lt 53,1%-ni. Kuna t-skoori täpsus ja saagis tõusevad korpuse mahu kasvamisega, siis saab öelda, et mida suurem korpus, seda paremad on t-skoori tulemused.

MI tulemused korpuse suuruse kasvades halvenevad: kui 5 miljonist sõnast koosnevast korpusest ühendverbide tuvastamisel on MI täpsus 2000 kandidaatpaari lõikes 14,1%, siis 170 miljoni sõna suuruse korpuse korral on täpsus 5,6%. Saagise väärtuse halveneb 16,2%-lt 6,4%-le. Eelnevalt tehtud oletus MI tulemuste halvenemise kohta peab paika ja korpuse suuruse kasvul on tugev negatiivne mõju MI tulemustele nii 100, 1000 kui ka 2000 kandidaatpaari lõikes.

Eespool tehtud oletust, et hii-ruut-statistiku tulemused korpuse mahu suurenedes paranevad, saab Tabeli 24 põhjal kinnitada: kui 5 miljonist sõnast koosneva korpuse 100 kandidaatpaari hulgas on täpsus 71,0% ja 2000 kandidaatpaari juures 31,6%, siis 170 miljoni sõna suuruse valimi korral on need näitajad vastavalt 79,0% ja 39,6%. Ka saagis paraneb 31,6%-lt 45,6%-ni. Järelikult parandab korpuse suurenemine hii-ruut-statistiku tulemusi ehk mida suurem on korpus, seda paremad on hii-ruut statistiku tulemused.

Log-tõepära funktsiooni tulemused paranevad korpuse kasvades 100 kandidaatpaari seas vähem, 2000 kandidaatpaari hulgas rohkem. Kui 5 miljonist sõnast ja 170 miljonist sõnast koosnevatest korpustest ühendverbide tuvastamisel on 100 kandidaatpaari lõikes log-tõepära funktsiooni täpsused ja saagised võrdsed, siis 20 miljoni sõna suuruse korpuse korral on tulemused mõnevõrra paranenud. 2000 kandidaatpaari seas on väiksema korpuse korral log-tõepära funktsiooni täpsus 38,8% ja saagis 44,7%, 170 miljonist sõnast koosneva korpuse puhul aga vastavalt 45,1% ja 52,0% Järelikult saab ümber lükata eespool esitatud väite, et korpuse mahu suurenemisel on log-tõepära funktsiooni tulemustele negatiivne mõju, ja öelda, et log-tõepära funktsiooni tulemused paranevad korpuse kasvades – väiksema arvu kandidaatpaaride seas küll vähem kui suurema arvu kandidaatpaaride hulgas.

MS-i 50 kõrgeima väärtusega ühendite loendeid võrreldes ja joonist 4 vaadeldes võis oletada, et MS-i tulemused paranevad korpuse suuruse kasvades. Tabelis 24 esitatud täpsuste võrdlemine kinnitab seda oletust. Kui võrrelda 5 miljoni suurust korpust 170 miljoni suuruse korpusega on vahed järgmised: MS-i täpsus suureneb 100 kandidaatpaari seas 80,0%-lt 83,0%-ni ja 2000 kandidaatpaari hulgas 35,9%-lt 40,2%-ni, saagis vastavalt 4,6%-lt 4,8%-ni ja 41,3%-lt 46,2%-ni. Seega MS-i tulemused paranevad korpuse mahu kasvamisega, mis tähendab, et mida suurem on korpus, seda paremad on MS-i tulemused. Kui eespool esitatud Wiechmanni (vt ptk 2.2.1.5) väidet kontrollida siinse töö tulemuste põhjal, siis ei saa nõustuda väitega, et valimi suuruse muutumine ei mõjuta MS-i tulemust nii suurel määral kui näiteks t-testi tulemusi, sest MS-i tulemused muutuvad rohkem kui t-testi omad. Samas hii-ruut-statistiku tulemusi mõjutab korpuse suuruse muutmine suuremal määral kui MS-i tulemusi.

Lihtsa sagedusloendi tulemusi võrreldes selgus, et korpuse kasvades selle tulemused ei parane ega halvene oluliselt. Tabeli 24 põhjal võib seda oletust kinnitada, sest 100 kandidaatpaari seas on lihtsa sagedusloendi täpsus ja saagis korpuse suurusest olenemata sama. 2000 kandidaatpaari hulgas on täpsus ja saagis kõige kõrgem 20 miljoni sõna suuruse valimi korral (vastavalt 41,9% ja 48,2%), 170 miljoni sõna suurusest korpusest ühendverbide tuvastamisel on lihtsa sageduse täpsus 41,4%, mis erineb 5 miljoni sõna suuruse valimile vastavast täpsusest (41,6%) vaid 0,2%. Niisamuti on lihtsa sagedusloendi saagis 170 miljoni sõna suuruse korpuse korral (47,7%) võrreldes 5 miljoni sõna suuruse korpusega (47,9%) kahanenud 0,2%. Järelikult lihtsa sagedusloendi tulemused korpuse mahu kasvades oluliselt ei muutu ehk korpuse suurusel pole mõju lihtsa sagedusloendi tulemustele.

3.3.2.8. Sümmeetriliste mõõdikute ja lihtsa sagedusloendi tulemuste kokkuvõte erineva suurusega korpuste põhjal

Tabeli 24 ja joonise 4 põhjal saab kokkuvõtlikult öelda, et kui korpuse maht suureneb, paranevad t-skoori, log-tõepära funktsiooni, hii-ruut-statistiku ja MS-i tulemused. Vastupidine mõju on korpuse mahu kasvamisel MI tulemustele. Lihtsa sagedusloendi tulemused jäävad korpuse kasvades peaaegu muutumatuks.

Korpuse suurusel ja vaadeldaval kandidaatpaaride arvul on mõju ka mõõdikute paremusjärjestusele. Parimaks mõõdikuks on olenemata korpuse suurusest või

vaadeldavate kandidaatpaaride arvust t-skoor ja halvimaks võrreldud statistikuks MI. Kui vaadelda 100 kandidaatpaari, siis on üldiselt MS parem lihtsast sagedusloendist, mille tulemused on omakorda paremad kui hii-ruut-statistikul. Siiski 20 miljoni sõna korpuse puhul on lihtsa sagedusloendi ja hii-ruut-statistiku tulemused võrdsed ja 170 miljoni sõna suuruses korpuses on hii-ruut-statistiku tulemused lihtsa sageduse omadest paremad. 2000 kandidaatpaari hulgas on üldjuhul log-tõepära efektiivsem kui lihtne sagedusloend, mille tulemused on omakorda paremad kui MS-i ja hii-ruut-statistiku omad. Vaid kõige väiksema korpuse korral on lihtne sagedusloend 2000 kandidaatpaari seas parem kui log-tõepära funktsioon.

3.2.3. Asümmeetriliste mõõdikute tulemused erineva suurusega korpuste põhjal

See peatükk kirjeldab asümmeetriliste mõõdikute tulemusi ühendverbide tuvastamisel erineva suurusega korpustest. Eesmärk on võrrelda, kuidas muutuvad asümmeetriliste mõõdikute tulemused, kui korpuse maht suureneb.

3.2.3.1. Tingliku tõenäosuse tulemused erineva suurusega korpuste põhjal

Tabelid 25 ja 26 esitavad kahanevalt 50 kõrgeima tingliku tõenäosuse väärtuse saanud ühendit 170 miljoni sõna suuruses ajakirjandustekstide korpuses. Tabelis 25 esitatud tingliku tõenäosuse väärtused väljendavad, kui suur on verbi esinemise tõenäosus, kui samas osalauses esineb ühendisse kuuluv adverb. Tabelis 26 esitatud tingliku tõenäosuse väärtused väljendavad aga seda, kui suure tõenäosusega esineb ühendisse kuuluv adverb, kui samas osalauses esineb ühendisse kuuluv verb.

Tabel 25. **50 kõrgeima CP(verb|adverb) väärtusega ühendit ajakirjandustekstides.**
Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; CP(2|1) – CP(verb|adverb) väärtus.

adverb	f1	verb	f2	sagedus	yv	CP(2|1)
pärale	841	jõudma	168218	773	TRUE	0,919
ülal	2362	pidama	702611	1733	TRUE	0,734
valmis	30479	olema	10739735	21917	FALSE	0,719
vahele	10777	jääma	360888	6249	TRUE	0,580
lahku	3331	minema	365466	1922	TRUE	0,577
peal	3084	olema	10739735	1759	FALSE	0,570

järel	2166	olema	10739735	1193	FALSE	0,551
kaasas	10455	olema	10739735	5754	FALSE	0,550
täis	16360	olema	10739735	8805	FALSE	0,538
tasa	2188	tegema	598768	1127	TRUE	0,515
üleval	5143	olema	10739735	2632	FALSE	0,512
takka	2	kihutama	9959	1	TRUE	0,500
takka	2	kiitma	28533	1	TRUE	0,500
takka	2	sundima	33781	1	TRUE	0,500
taha	1095	jääma	360888	488	TRUE	0,446
kallale	3189	tungima	8817	1385	FALSE	0,434
pihta	3769	saama	1057514	1512	TRUE	0,401
pihta	3769	hakkama	266118	1463	TRUE	0,388
vastu	76337	võtma	322185	28435	TRUE	0,372
koos	27259	olema	10739735	9566	FALSE	0,351
alt	4217	vedama	24469	1455	TRUE	0,345
ilma	21833	jääma	360888	7505	TRUE	0,344
valla	3269	olema	10739735	1085	FALSE	0,332
taga	7767	olema	10739735	2457	FALSE	0,316
kõrvalt	2316	vaatama	114220	722	TRUE	0,312
otsa	13292	saama	1057514	3725	TRUE	0,280
ligi	44960	olema	10739735	12346	FALSE	0,275
kinni	60871	pidama	702611	16411	TRUE	0,270
pealt	5394	vaatama	114220	1451	TRUE	0,269
ülal	2362	olema	10739735	624	FALSE	0,264
taga	7767	ajama	46586	2019	TRUE	0,260
üle	134583	olema	10739735	34501	FALSE	0,256
ühte	2132	olema	10739735	546	FALSE	0,256
esile	15517	tõstma	46220	3920	TRUE	0,253
läbi	91173	olema	10739735	23009	FALSE	0,252
otsa	13292	sõitma	115057	3339	TRUE	0,251
ilma	21833	olema	10739735	5387	FALSE	0,247
esile	15517	tooma	142990	3732	TRUE	0,241
lahti	34458	olema	10739735	8062	FALSE	0,234
ette	114045	nägema	200795	26025	TRUE	0,228
ette	114045	olema	10739735	25984	FALSE	0,228
kokku	150217	olema	10739735	33988	FALSE	0,226
peale	33718	olema	10739735	7494	FALSE	0,222
kaasa	75124	tooma	142990	16097	TRUE	0,214
üleval	5143	pidama	702611	1082	TRUE	0,210
eemale	8654	jääma	360888	1801	TRUE	0,208
ümber	26581	olema	10739735	5516	FALSE	0,208

alla	59223	kirjutama	132506	12218	TRUE	0,206
kaasas	10455	käima	172564	2124	TRUE	0,203
kõrvalt	2316	olema	10739735	470	FALSE	0,203

Tabel 26. **50 kõrgeima CP(adverb|verb) väärtusega ühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; CP(1|2) – CP(adverb|verb) väärtus.

adverb	**f1**	**verb**	**f2**	**sagedus**	**yv**	**CP(1|2)**
sekka	1124	jõratama	1	1	FALSE	1
sekka	1124	mahakooruma	1	1	FALSE	1
sekka	1124	taas-taas-taasavaldama	1	1	FALSE	1
ühte	2132	kanseldama	1	1	FALSE	1
järel	2166	mitte-kunagi-armastama	1	1	FALSE	1
tagant	2506	turgima	1	1	FALSE	1
kallale	3189	õppi-matma	1	1	FALSE	1
kallale	3189	tromama	1	1	FALSE	1
valla	3269	portesteerima	1	1	FALSE	1
alt	4217	delegeeruma	1	1	FALSE	1
alt	4217	tsirukleerima	1	1	FALSE	1
üleval	5143	walitsema	1	1	FALSE	1
ümber	26581	rahvustuma	5	5	TRUE	1
taga	7767	garaaxima	1	1	FALSE	1
taga	7767	pusletama	1	1	FALSE	1
taga	7767	nõduma	1	1	FALSE	1
eemale	8654	seminaritsema	1	1	FALSE	1
eemale	8654	otsas/muigama	1	1	FALSE	1
eemale	8654	jää-minema	1	1	FALSE	1
eemale	8654	hoidsima	1	1	FALSE	1
eemale	8654	põnklema	1	1	FALSE	1
eemale	8654	neljutama	1	1	FALSE	1
eemale	8654	timpima	1	1	FALSE	1
püsti	9341	ballastima	1	1	FALSE	1
laiali	9712	kon-marineerima	1	1	FALSE	1
laiali	9712	toosistama	1	1	FALSE	1
laiali	9712	lammutatama	1	1	FALSE	1
laiali	9712	peks-matma	1	1	FALSE	1
ringi	23320	kosserdama	2	2	FALSE	1
ringi	23320	höörutama	2	2	FALSE	1
otsa	13292	pilguma	1	1	FALSE	1

otsa	13292	lõppetama	1	1	FALSE	1
otsa	13292	omima	1	1	FALSE	1
mööda	13752	pargi/lõikama	1	1	FALSE	1
mööda	13752	melodeklameerima	1	1	FALSE	1
mööda	13752	hargi/roitma	1	1	FALSE	1
mööda	13752	rümama	1	1	FALSE	1
esile	15517	kutssuma	1	1	FALSE	1
esile	15517	eakaas-laskma	1	1	FALSE	1
kõrvale	15717	tõtta-tõttama	1	1	FALSE	1
kõrvale	15717	dopima	1	1	FALSE	1
kõrvale	15717	mõrvuma	1	1	FALSE	1
kõrvale	15717	emotsionaalse-ärritama	1	1	FALSE	1
kõrvale	15717	mitte-olenema	1	1	FALSE	1
kõrvale	15717	marrssima	1	1	FALSE	1
kõrvale	15717	hiillima	1	1	FALSE	1
täis	16360	jürima	1	1	FALSE	1
täis	16360	kriipsima	1	1	FALSE	1
peale	33718	tappa-tükeldama	2	2	FALSE	1
järele	17756	kikima	1	1	FALSE	1

50-st kõrgeima CP(verb|adverb) väärtuse saanud ühendist on 28 õiged ühendverbid. Võrreldes Tasakaalus korpuse ajakirjandustekstide 50 kõrgeima CP(verb|adverb) väärtusega (vt tabel 11) tuvastab CP(verb|adverb) 170 miljonist sõnast koosnevast korpusest kolm õiget ühendverbi rohkem. Seega korpuse suuruse kasvamisega paranevad CP(verb|adverb) tulemused. 170 miljoni sõna suuruse ajakirjanduskorpuse 50 kõrgeima CP(adverb|verb) väärtusega ühendi seas on üks õige ühendverb. Ka Tasakaalus korpuse 50-st kõrgeima CP(adverb|verb) väärtusega ühendist on üks õige ühendverb (vt Tabel 12). Järelikult CP(adverb|verb) tulemused ei muutu korpuse mahu kasvamisega.

Tingliku tõenäosuse tulemuste seosed korpuse mahu suurenemisega sõltuvad sellest, missugust seost vaadelda: kui tegemist on verbi esinemise sõltumisega adverbi esinemisest, siis tingliku tõenäosuse tulemused kasvavad korpuse mahu suurenemisega, kuid vastupidise seose puhul tulemused ei muutu ja korpuse suurus ei mõjuta tingliku tõenäosuse tulemusi.

3.2.3.2. ΔP tulemused erineva suurusega korpuste põhjal

Tabelid 27 ja 28 esitavad kahanevalt 50 kõrgeima ΔP väärtuse saanud ühendit 170 miljoni sõna suuruses ajakirjandustekstide korpuses. Tabelis 27 väljendavad ΔP väärtused, kui

tugevalt sõltub verbi esinemine adverbi esinemisest samas osalauses, tabelis 28 aga seda, kui tugevalt sõltub ühendisse kuuluva adverbi esinemine verbi esinemist samas osalauses.

Tabel 27. **50 kõrgeima ΔP(verb|adverb) väärtusega ühendit ajakirjandustekstides.**
Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; ΔP(2|1) – ΔP(verb|adverb) väärtus.

adverb	**f1**	**verb**	**f2**	**sagedus**	**yv**	**ΔP(2|1)**
pärale	841	jõudma	168218	773	TRUE	0,912
ülal	2362	pidama	702611	1733	TRUE	0,705
vahele	10777	jääma	360888	6249	TRUE	0,565
lahku	3331	minema	365466	1922	TRUE	0,562
takka	2	kihutama	9959	1	TRUE	0,500
takka	2	kiitma	28533	1	TRUE	0,499
takka	2	sundima	33781	1	TRUE	0,499
tasa	2188	tegema	598768	1127	TRUE	0,491
kallale	3189	tungima	8817	1385	FALSE	0,434
taha	1095	jääma	360888	488	TRUE	0,431
pihta	3769	hakkama	266118	1463	TRUE	0,377
vastu	76337	võtma	322185	28435	TRUE	0,360
pihta	3769	saama	1057514	1512	TRUE	0,358
alt	4217	vedama	24469	1455	TRUE	0,344
ilma	21833	jääma	360888	7505	TRUE	0,329
kõrvalt	2316	vaatama	114220	722	TRUE	0,307
valmis	30479	olema	10739735	21917	FALSE	0,278
pealt	5394	vaatama	114220	1451	TRUE	0,264
taga	7767	ajama	46586	2019	TRUE	0,258
esile	15517	tõstma	46220	3920	TRUE	0,251
otsa	13292	sõitma	115057	3339	TRUE	0,247
kinni	60871	pidama	702611	16411	TRUE	0,241
otsa	13292	saama	1057514	3725	TRUE	0,237
esile	15517	tooma	142990	3732	TRUE	0,235
ette	114045	nägema	200795	26025	TRUE	0,221
kaasa	75124	tooma	142990	16097	TRUE	0,209
alla	59223	kirjutama	132506	12218	TRUE	0,201
peal	3084	hoidma	76157	625	FALSE	0,200
kaasas	10455	käima	172564	2124	TRUE	0,196
esile	15517	kutsuma	63938	3074	TRUE	0,196
eemale	8654	jääma	360888	1801	TRUE	0,193

pealt	5394	nägema	200795	1081	TRUE	0,192
üleval	5143	pidama	702611	1082	TRUE	0,182
püsti	9341	tõusma	61765	1688	FALSE	0,178
mööda	13752	minema	365466	2604	TRUE	0,174
kõrvale	15717	jääma	360888	2928	TRUE	0,172
üleval	5143	hoidma	76157	895	FALSE	0,171
laiali	9712	saatma	84129	1681	FALSE	0,170
külge	2997	jääma	360888	532	TRUE	0,163
lahku	3331	lööma	53563	543	TRUE	0,161
alla	59223	jääma	360888	10308	TRUE	0,160
eemale	8654	hoidma	76157	1390	FALSE	0,158
juurde	24763	tulema	567778	4295	TRUE	0,150
kallale	3189	minema	365466	525	FALSE	0,150
taga	7767	otsima	55335	1168	TRUE	0,148
püsti	9341	panema	178482	1441	FALSE	0,147
kõrvale	15717	jätma	101181	2321	TRUE	0,144
läbi	91173	viima	116301	12750	TRUE	0,136
pealt	5394	kuulama	26953	729	TRUE	0,134
vahele	10777	jätma	101181	1483	TRUE	0,134

Tabel 28. **50 kõrgeima ΔP(adverb|verb) väärtusega ühendit ajakirjandustekstides.**

Lühendid: f1 – adverbi sagedus korpuses, f2 – verbi sagedus korpuses; sagedus – ühendi koosesinemissagedus samas osalauses; yv – ühendverbi väärtus: TRUE = on õige ühendverb, FALSE = ei ole õige ühendverb; ΔP(1|2) – ΔP(adverb|verb) väärtus.

adverb	**f1**	**verb**	**f2**	**sagedus**	**yv**	**ΔP(1|2)**
sekka	1124	jõratama	1	1	FALSE	1
sekka	1124	mahakooruma	1	1	FALSE	1
sekka	1124	taas-taas-taasavaldama	1	1	FALSE	1
ühte	2132	kanseldama	1	1	FALSE	1
järel	2166	mitte-kunagi-armastama	1	1	FALSE	1
tagant	2506	turgima	1	1	FALSE	1
kallale	3189	õppi-matma	1	1	FALSE	1
kallale	3189	tromama	1	1	FALSE	1
valla	3269	portesteerima	1	1	FALSE	1
alt	4217	delegeeruma	1	1	FALSE	1
alt	4217	tsirukleerima	1	1	FALSE	1
üleval	5143	walitsema	1	1	FALSE	1
taga	7767	garaaxima	1	1	FALSE	1
taga	7767	pusletama	1	1	FALSE	1
taga	7767	nõduma	1	1	FALSE	1

eemale	8654	seminaritsema	1	1	FALSE	1
eemale	8654	otsas/muigama	1	1	FALSE	1
eemale	8654	jää-minema	1	1	FALSE	1
eemale	8654	hoidsima	1	1	FALSE	1
eemale	8654	põnklema	1	1	FALSE	1
eemale	8654	neljutama	1	1	FALSE	1
eemale	8654	timpima	1	1	FALSE	1
püsti	9341	ballastima	1	1	FALSE	1
laiali	9712	kon-marineerima	1	1	FALSE	1
laiali	9712	toosistama	1	1	FALSE	1
laiali	9712	lammutatama	1	1	FALSE	1
laiali	9712	peks-matma	1	1	FALSE	1
otsa	13292	pilguma	1	1	FALSE	0,999
otsa	13292	lõppetama	1	1	FALSE	0,999
otsa	13292	omima	1	1	FALSE	0,999
mööda	13752	pargi/lõikama	1	1	FALSE	0,999
mööda	13752	melodeklameerima	1	1	FALSE	0,999
mööda	13752	hargi/roitma	1	1	FALSE	0,999
mööda	13752	rümama	1	1	FALSE	0,999
esile	15517	kutssuma	1	1	FALSE	0,999
esile	15517	eakaas-laskma	1	1	FALSE	0,999
kõrvale	15717	tõtta-tõttama	1	1	FALSE	0,999
kõrvale	15717	dopima	1	1	FALSE	0,999
kõrvale	15717	mõrvuma	1	1	FALSE	0,999
kõrvale	15717	emotsionaalse-ärritama	1	1	FALSE	0,999
kõrvale	15717	mitte-olenema	1	1	FALSE	0,999
kõrvale	15717	marrssima	1	1	FALSE	0,999
kõrvale	15717	hiillima	1	1	FALSE	0,999
täis	16360	jürima	1	1	FALSE	0,999
täis	16360	kriipsima	1	1	FALSE	0,999
järele	17756	kikima	1	1	FALSE	0,999
järele	17756	analüseerima	1	1	FALSE	0,999
ilma	21833	etiketistama	1	1	FALSE	0,999
ilma	21833	verd-jõudma	1	1	FALSE	0,999
ilma	21833	äirima	1	1	FALSE	0,999

170 miljoni sõna suuruse ajakirjandustekstide korpuse põhjal leitud 50-st kõrgeima ΔP(verb|adverb) väärtusega ühendist on 41 õiged ühendverbid. Võrreldes Tasakaalus korpuse põhjal tehtud 50 kõrgeima ΔP(verb|adverb) väärtusega ühendi loeteluga (vt tabel

13), on õigete ühendverbide arv kasvanud 7 võrra. Seega korpuse mahu suurenemine parandab ΔP(verb|adverb) tulemusi.

Tabelis 28 ei ole ühtegi sellist ühendit, mis kuuluks EKSS-i loendisse. Tasakaalus korpuse ajakirjandustekstide põhjal tehtud 50 kõrgeima väärtusega ühendi nimekirjas on õigeid ühendverbe 1 (vt tabel 14). Järelikult korpuse mahu suurenedes ΔP(adverb|verb) tulemused halvenevad veelgi. Samas on see paratamatu, sest korpuse kasvades tuleb juurde ka selliseid ühendeid, mis esinevad korpuses vaid ühe korra ja mille verbiline komponent on samuti harvaesinev ning seega kasvab selliste ühendite arv, mis saavad kõrge ΔP(adverb|verb) väärtuse.

3.2.3.3. Kokkuvõte asümmeetriliste mõõdikute tulemustest erineva suurusega korpuste põhjal

Selleks et vaadelda, kuidas korpuse mahu suurenemine mõjutab asümmeetriliste statistikute tulemusi, saab vaadelda nende tuvastatud õigete ühendverbide arvu. Täpsema ülevaate korpuse mahu suurenemise mõjust asümmeetriliste statistikute tulemustele saab tabelist 29, kus on esitatud asümmeetriliste mõõdikute tuvastatud õigete ühendverbide arvud erineva suurusega ajakirjandustekstide korpustes.

Tabel 29. **Asümmeetriliste statistikute tuvastatud õigete ühendverbide hulk erineva suurusega ajakirjanduskorpustes.**

sõnade arv korpuses	CP(verb|adverb)	CP(adverb|verb)	ΔP(verb|adverb)	ΔP(adverb|verb)
5 000 000	25	1	36	1
10 000 000	25	1	39	1
20 000 000	26	0	38	0
70 000 000	26	0	40	0
170 000 000	28	1	41	0

Kui vaadelda seost, kuidas verbi esinemine sõltub adverbi esinemisest samas osalauses, siis nii tingliku tõenäosuse kui ka ΔP tulemused korpuse kasvades paranevad: kui 5 miljoni sõna suurusest korpusest tuvastas tinglik tõenäosus 25 ja ΔP 36 õiget ühendverbi, siis 170 miljoni sõna suuruse korpuse põhjal on 50 kõrgeima väärtuse saanud ühendi seas need numbrid vastavalt 28 ja 41. Vastupidist seost uurides aga tulemused korpuse mahu suurendes ei muutu või halvenevad. Seega korpuse mahu kasvu mõju oleneb asümmeetriliste statistikute puhul sellest, missugust seost sõnade vahel vaadelda.

Ka erinevate korpuse suuruste kohta kehtivad juba eespool (vt ptk 3.1.3.3), tekstiliikide võrdlemisel tehtud järeldused: CP(adverb|verb) ja ΔP(adverb|verb) ei tuvasta palju õigeid ühendverbe, kuid sobivad harva esinevate ühendite tuvastamiseks ja nad sisaldavad informatsiooni ühendite komponentide seoste kohta.

3.2.4. Sümmeetriliste ja asümmeetriliste mõõdikute tulemuste võrdlus

Selleks et teada saada, kas ja kuidas sümmeetriliste ja asümmeetriliste statistikute tulemused ühendverbide tuvastamisel sõltuvad korpuse suuruse kasvamisest, ning kas asümmeetrilised statistikute tulemused erinevad sümmeetriliste statistikute tulemustest, võrdlen tuvastatud õigete ühendverbide loendeid. Aluseks võtan eespool esitatud loendid, kus on esitatud iga mõõdiku 50 kõrgeima väärtusega ühendit ning loen kokku mõõdikute tuvastatud erinevad õiged ühendverbid. Tabelis 30 on esitatud sümmeetriliste ja asümmeetriliste statistikute tuvastatud õigete ühendverbide arvud erineva suurusega korpustest, kui iga mõõdiku puhul võtta arvesse 50 kõige kõrgema statistiku väärtuse saanud ühendit. Samuti on tabelis 30 esitatud nende õigete ühendverbide arvud, mida sümmeetrilised statistikud tuvastavad, kuid asümmeetrilised ei tuvasta, ning nende õigete ühendverbide arvud, mida asümmeetrilised tuvastavad, kuid sümmeetrilised ei tuvasta.

Tabel 30. **Sümmeetriliste ja asümmeetriliste statistikute tuvastatud õigete ühendverbide hulk erineva suurusega ajakirjanduskorpustes.**

sõnade arv korpuses	sümmeetrilised	asümmeetrilised	ainult sümmeetrilised	ainult asümmeetrilised
5 000 000	83	37	69	23
10 000 000	82	40	67	25
20 000 000	86	38	72	24
70 000 000	86	40	70	24
170 000 000	85	42	70	27

5 miljoni sõna suurusest ajakirjanduskorpusest tuvastavad asümmeetrilised mõõdikud 23 sellist ühendverbi, mida sümmeetrilised ei tuvasta. 170 miljoni sõna suurusest korpusest tuvastavad sümmeetrilised statistikud 85 erinevat õiget ühendverbi, asümmeetrilised aga 42, millest 27 on sellised, mida sümmeetrilised ei tuvasta. Korpuse kasvades tuvastavad nii sümmeetrilised kui ka asümmeetrilised mõõdikud rohkem õigeid ühendverbe ehk nende tulemuslikkus kasvab. Korpuse mahu suurenedes kasvab ka nende õigete ühendverbide arv, mida asümmeetrilised erinevalt sümmeetrilistest tuvastavad.

Seega olenemata korpuse suurusest tasub ühendverbide tuvastamisel asümmeetrilisi mõõdikuid kasutada, sest kaks asümmeetrilist mõõdikut suudavad erinevas suuruses korpusest tuvastada selliseid ühendverbe, mida viis sümmeetrilist ja lihte sagedusloend ei suuda.

Ka erinevate suurusega korpuste kohta kehtivad juba eespool (vt ptk 3.1.4), tekstiliikide võrdlemisel tehtud järeldused: teistest oluliselt paremat mõõdikut ei ole ja mõistlik on erinevat liiki statistikuid omavahel kombineerida.

3.3. Sümmeetriliste ja asümmeetriliste mõõdikute tulemuste võrdlus

Eespool (vt ptk 3.1.4 ja 3.2.4) selgus, et püsiühendite tuvastamiseks tasub laialt kasutatud sümmeetriliste mõõdikute kõrval rakendada ka psühholoogiliselt paremini põhjendatud asümmeetrilisi statistikuid. See peatükk võrdleb sümmeetriliste ja asümmeetriliste mõõdikute tulemusi ja eesmärk on välja selgitada, kas asümmeetrilised statistikud on tulemuslikud ka suurema arvu kandidaatpaaride korral kui 50, ning kas ja mida täpsemalt nende rakendamine lõpptulemusele juurde annab.

Eesmärgi saavutamiseks kasutan 170 miljonist sõnast koosnevat ajakirjanduskorpust ja võrdlen sümmeetrilisi statistikuid – t-skoori, MI-d, hii-ruut-statistikut, log-tõepära funktsiooni ja MS-i – asümmeetrilise ΔP-ga. Asümmeetriliste statistikutest olen valinud paremate tulemuste põhjal ΔP (vt ptk 3.1.3.3 ja 3.2.3.3).

Sümmeetriliste mõõdikute ja asümmeetrilise ΔP tulemuste võrdlemiseks võtan arvesse need ühendid, mille ΔP-de väärtuste vahe (ΔP(verb|adverb)–ΔP(adverb|verb)) on suurem või väiksem kui 0 ehk ühendid, mida võib pidada asümmeetrilisteks. Selliseid ühendeid on ajakirjanduskorpuses 3314, millest 364 on õiged ühendverbid. ΔP võrdlemiseks sümmeetriliste statistikutega arvestan iga sümmeetrilise statistiku 3314 kõrgeima väärtuse saanud ühendit, mille seast eraldan õiged ühendverbid. Kõikide sümmeetriliste statistikute tuvastatud õigetest ühendverbidest moodustan ühise loendi, kus iga ühendverbi esineb vaid korra. ΔP hindamiseks võrdlen sümmeetriliste statistikute tuvastatud õigete ühendverbide loendit ΔP tuvastatud õigete ühendverbide nimekirjaga.

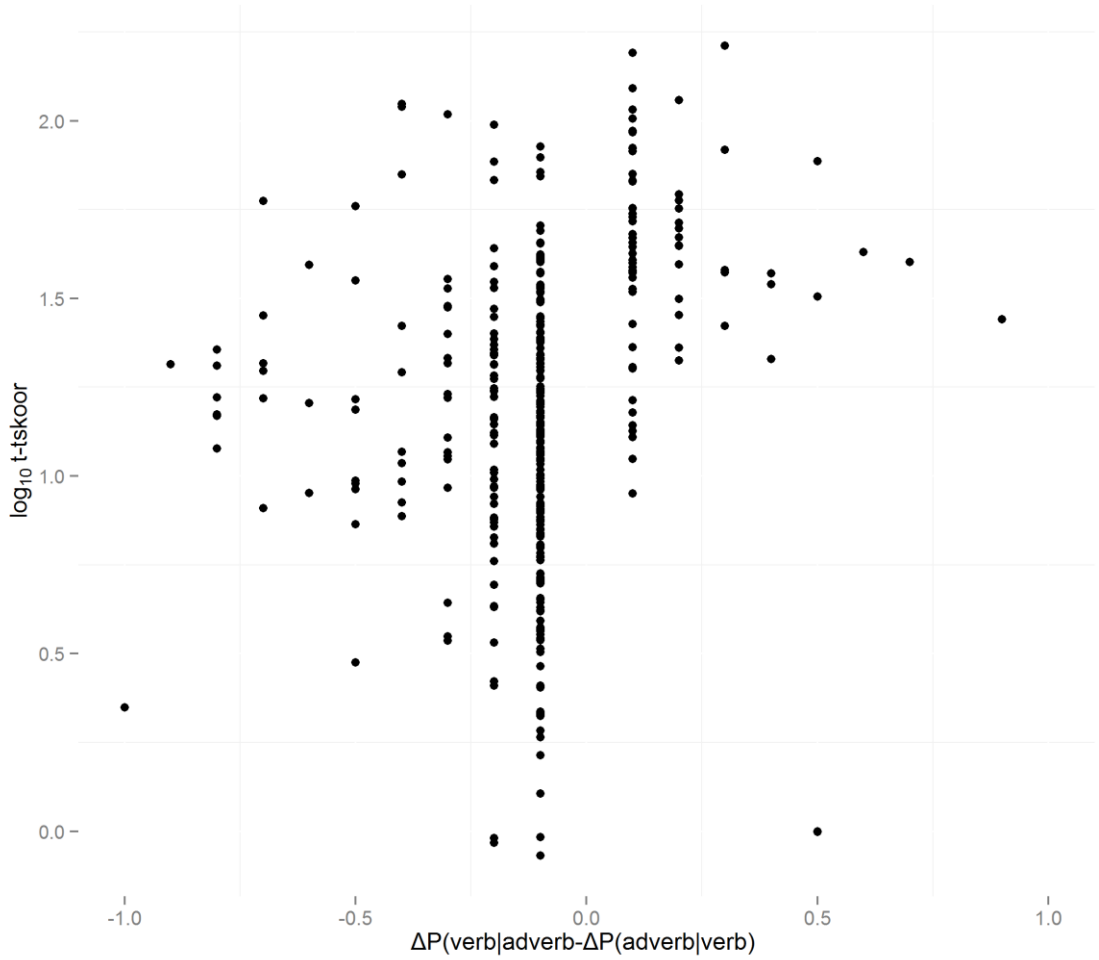
Selgub, et sümmeetrilised statistikud tuvastavad kokku 1249 õiget ühendverbi. ΔP ja sümmeetriliste mõõdikute tuvastatud õigete ühendverbide loendite võrdlemisel selgus, et ΔP tuvastatud 364 ühendverbist on 14 sellist, mida sümmeetrilised statistikud ei tuvasta. Kuigi sümmeetrilised mõõdikud tuvastavad kokku rohkem õigeid ühendverbe,

tuvastab asümmeetriline ΔP üksi selliseid ühendverbe, mida sümmeetrilised isegi koos ei tuvasta ja ΔP rakendamine ühendverbide tuvastamisel on põhjendatud.

Kõige efektiivsemaks sümmeetriliseks statistikuks osutub eesti keele ühendverbide automaatsel tuvastamisel t-skoor, mille 3314 kõrgeima väärtuse saanud ühendi seas on 1082 õiget ühendverbi. Nende ühendverbide võrdlemisel ΔP tuvastatud ühendverbidega selgub, et ΔP õigete ühendverbide hulgas on 34 sellist ühendverbi, mida t-skoor tuvastada ei suuda. t-skoor tuvastab küll rohkem õigeid ühendverbe kui ΔP ja on edukam eesti keele ühendverbide tuvastamisel, ent ΔP tulemused lisavad infot selle kohta, kumb sõna kollokatsioonis teisest rohkem sõltub.

Kõrgeima positiivse ΔP kahe väärtuse vahe saanud ühendverb on *pärale jõudma*. Selles ühendverbis on verb *jõudma* palju rohkem ennustatav adverbi *pärale* abil kui vastupidi ehk kui osalauses esineb adverb *pärale*, siis esineb seal suure tõenäosusega ka verb *jõudma* aga kui osalauses esineb verb *jõudma*, siis adverbi *pärale* esinemine nii tõenäoline ei ole. Põhjus peitub selles, et *pärale* on adverb, mis üldjuhul esinebki ühendverbi *pärale jõudma* koosseisus. Madalaim ΔP väärtuste vahe on ühendverbil *ümber rahvustuma*, milles on adverb *ümber* palju rohkem ennustatav verbi *rahvustuma* abil kui vastupidi ehk kui osalauses esineb verb *rahvustuma*, siis esineb seal suure tõenäosusega adverb *ümber* aga kui osalauses esineb adverb *ümber*, siis verbi *rahvustuma* esinemine nii tõenäoline ei ole.

Seda, et sümmeetrilised mõõdikud ei sisalda infot ühendite asümmeetrilisuse kohta, kinnitab joonis 5, kus on esitatud ΔP tuvastatud õigete ühendverbide jaotus nende t-skoori ja ΔP(verb|adverb) – ΔP(adverb|verb) väärtuste järgi. Joonise x-teljel on ΔP kahe väärtuse vahe ning y-teljel t-skoori väärtused. t-skoori väärtused on logaritmitud alusel 10.



Joonis 5. **ΔP tuvastatud õigete ühendverbide jaotus ΔP(verb|adverb) – ΔP(adverb|verb) ja t-skoori väärtuste järgi.**

Jooniselt 5 selgub, et kõrge t-skoori väärtusega ühendeid leidub igasuguse ΔP väärtusega ühendite seas, mis tähendab, et t-skoor ja üldistatult ka teised sümmeetrilised statistikud omistavad ühenditele kõrgeid kahesuunalisi väärtuseid arvestamata seda, kumb sõna teisest sõltub. Asümmeetrilise ΔP abil saab aga tuvastada ühendeid, mille komponentide vahel on asümmeetriline seos.

Jooniselt 5 on näha, et rohkem on negatiivse ΔP vahega ühendeid ja järelikult on andmestikus rohkem ühendeid, kus verb on paremini ennustatav adverbi järgi kui vastupidi. Samas t-skoor omistab sagedamini kõrgema väärtuse just nendele ühenditele, kus adverb on paremini ennustatav verbi esinemise järgi kui vastupidi, ning seega tuvastab paremini üht tüüpi ühendverbe.

Kokkuvõttes võib öelda, et ΔP-d ja üldistatult teisi asümmeetrilisi mõõdikuid on mõistlik sümmeetriliste mõõdikute kõrval kollokatsioonide tuvastamisel rakendada, sest nad tuvastavad ühendeid, mida sümmeetrilised mõõdikud ei tuvasta, ning sisaldavad lisainfot ühendite asümmeetrilisuse kohta. Katsetulemused osutavad, et asümmeetrilistele statistikutele peaks rohkem tähelepanu pöörama ning kindlasti peaks nende omadusi tulevikus kollokatsioonide tuvastamise kontekstis rohkem uurima.

3.4. Tulemuste võrdlus teiste keelte sarnaste katsetega

Kollokatsioonide tuvastamise katseid on kõige rohkem tehtud inglise, saksa ja prantsuse keelte põhjal. Vähem on sõnadevahelise seose tugevuse mõõdikuid testitud muude keelte andmestike peal. (Seretan 2011: 49–58) Hoolimata arvukatest katsetest, pole erinevad autorid jõudnud üldiste ja lõplike järeldusteni (Evert 2008: 37). Järgnevalt annan ülevaate mõnest meetodeid kõrvutavast tööst, mille tulemusi võrdlen siinses töös kirjeldatud katsete tulemustega.

Daille (1996) tuvastas prantsuse keelest mitmest nimisõnast koosnevad termineid ning selle ülesande lahendamiseks osutus parimaks log-tõepära funktsioon. Siinses töös log-tõepära funktsioon küll kõige kõrgemaid täpsuse väärtuseid ei saavutanud, kuid seda võib heade tulemuste põhjal pidada sobivaks meetodiks ka eesti keele ühendverbide tuvastamisel.

Krenn ja Evert (2001) võrdlesid sümmeetrilisi sõnadevahelise seose tugevuse mõõdikuid saksa keele eessõna ja verbi ühendite tuvastamisel ning võrdlesid t-testi, log-tõepära funktsiooni, hii-ruut-statistikut ja MI-d lihtsa sagedusloendi tulemusega. Selles katses saavutas kõige kõrgemaid täpsuse väärtuseid t-skoor, kuid autorid leidsid, et ükski võrreldud mõõdik polnud tähelepanuväärselt parem kui lihtne sagedusloend. Samas laialt kasutatava log-tõepära funktsiooni tulemused on saksa keelest eessõna ja verbi ühendite tuvastamisel märgatavalt halvemad kui lihtsa sagedusloendi omad. Kui kõrvutada neid tulemusi selles töös kirjeldatud katse tulemustega, siis mõlemas katses saavutas t-skoor parima tulemuse, kuid tulemused polnud väga palju paremad lihtsa sagedusloendi omadest. Vastupidiselt saksa keele põhjal tehtud katsega, polnud eesti keele ühendverbide tuvastamisel log-tõepära funktsiooni tulemused märkimisväärselt halvemad kui lihtsa sagedusloendi tulemused. Sellised erinevused võivad viidata varem korduvalt

väljatoodud tähelepanekutele (nt Krenn 2000; Ramisch jt 2008), et mõõdiku tulemuslikkust mõjutab ka tuvastatava püsiühendi liik. Krenn ja Evert leidsid, et tulemused ei sõltu oluliselt tekstitüübist, kuid siinses töös selgus, et tekstiliigil on mõju statistikute tulemustele.

Kis jt (2003) tegid katse selliste ungari keele mitmesõnaliste ühendite tuvastamiseks, mis koosnesid verbist, nimisõnast ja nimisõna käändetunnusest. Iga trigramm teisendati kaheks bigrammiks: verb-nimisõna ja nimisõna-käändetunnus. Selliste bigrammide tuvastamiseks rakendasid autorid kahte statistilist meetodit – log-tõepära funktsiooni ja esilduvuse mõõdikut (*salience*). Lisaks vaatlesid ka lihtsa sagedusloendi tulemusi. Meetodite valik põhines varasemate katsete tulemustel: valiti välja meetodid, mis hollandi keele kollokatsioonide tuvastamisel saavutasid parima täpsuse koos parima saagisega. Selgus, et 100 kõrgeima väärtusega kandidaatpaari seas oli parim meetod esilduvus, kuid ka log-tõepära funktsiooni tulemused olid märkimisväärselt head. Mõlemad meetodid olid paremad ka algtaseme väärtusest. Lihtsa sagedusloendi tulemused olid selgelt halvemad teistest meetoditest. Siinses töös kirjeldatud katsete põhjal selgus samuti, et 100 kõrgeima väärtusega kandidaatpaari võrdluses on log-tõepära funktsiooni tulemused paremad kui lihtsa sagedusloendi omad. Samuti leidsid Kis jt, et ungari keele peal saab kasutada samasid statistilisi meetodeid, mida hollandi keele peal, kuid mõned kollokatsioonitüübid nõuavad teistsuguseid tuvastamise meetodeid, mis omakorda kinnitab taas, et pole olemas universaalset meetodit kollokatsioonide tuvastamiseks kõikidest keeltest ja seetõttu on katsete kordamine erinevates keeltes oluline.

Pecina ja Schlesinger (2006) tuvastasid tšehhi keele kahesõnalisi kollokatsioone – idiomaatilisi väljendeid, tehnilisi termineid, tugiverbiühendeid, nimesid ja tüüpfraase (*stock phrases)* – ja võrdlesid 82 erinevat sõnadevahelise seose tugevuse mõõdikut, mille hulka kuulusid ka t-skoor, log-tõepära funktsioon, hii-ruut-statistik ja tinglik tõenäosus. Tšehhi keele andmestiku peal oli parimate meetodite hulgas hii-ruut-statistik, mille tulemused on selles töös tehtud katsete põhjal selgelt halvemad kui t-skoori ja log-tõepära funktsiooni tulemused.

Michelbacher jt (2011) võrdlesid inglise keele kahesõnalisi (kas adjektiivist ja nimisõnast või kahest nimisõnast koosnevaid) süntagmaatilisi kombinatsioone nii

korpus- kui ka eksperimentaalse andmestiku põhjal ning demonstreerisid, et katseisikute produtseeritud süntagmaatiliste seoste asümmeetrilisus korreleerub tugevalt tingliku tõenäosusega. Nende töö baseerus neljal seosetüübil: paradigmaatiline – süntagmaatiline ja sümmeetriline – asümmeetriline, näiteks seos sõnade *halb* (*bad*) ja *hea* (*good*) vahel on paradigmaatiline ja sümmeetriline ning sõnade *jõulud* (*Christmas*) ja *kaunistused* (*decorations*) vahel aga süntagmaatiline ja asümmeetriline. Autorid leidsid, et tinglik tõenäosus sobib tugevalt asümmeetriliste seoste tuvastamiseks, kuid ei suuda tuvastada sümmeetrilisi seoseid. Michelbacheri jt katsetest selgus ka, et suur osa uuritavatest kombinatsioonidest olid asümmeetrilised ja paremalt poolt ennustatav asümmeetria on rohkem levinud kui vasakult poolt ennustatav asümmeetria. Paremalt poolt ennustatav asümmeetria tähendab, et eelnev sõna ennustab järgmist sõna suurema tõenäosusega kui järgnev sõna eelnevat, vasakult poolt ennustatav asümmeetria aga, et järgnev sõna ennustab eelnevat suurema tõenäosusega kui eelnev järgnevat sõna. Paremalt poolt ennustatava asümmeetria on autorite järgi levinum, sest eelnev sõna on oluline faktor järgneva sõna ennustamisel. Nii nagu Michelbacheri jt katses tinglik tõenäosus, tuvastavad ka siinses töös mõlemad asümmeetrilised mõõdikud – tinglik tõenäosus ja ΔP – paremini asümmeetrilisi kui sümmeetrilisi seoseid. Näiteks 50 kõrgeima statistiku väärtuse saanud kandidaatpaari seas (vt nt tabelid 25 ja 26 või tabelid 27 ja 28) ei ole ühendeid, mille puhul on verb ennustatav adverbi järgi sama tõenäosusega, mis adverb verbi järgi. Nagu ka inglise keele adjektiivi-nimisõna ja nimisõna-nimisõna kombinatsioonide hulgas on ka eesti keele ühendverbides rohkem levinud paremalt poolt ennustatav asümmeetria kui vasakult poolt ennustatav asümmeetria ehk rohkem on ühendeid, kus adverb ennustab paremini verbi esinemist kui verb adverbi esinemist. Näiteks kuuluvad sellesse hulka ühendid *pärale jõudma*, *ülal pidama* ja *lahku minema*, mille puhul on ühendi verbiline komponent paremini ennustatav adverbi esinemise järgi kui vastupidi. See on ka loogiline, sest need afiksaaladverbid esinevad harva kui üldse mingis muus kontekstis.

Kokkuvõtvalt on eesti keele ühendverbide tuvastamise katse tulemused sarnased saksa keele eessõnast ja verbist koosnevate ühendite tuvastamise tulemustega – mõlemas katses esineb kõige paremini t-skoor. Samuti kinnitas siinne katse Michelbacheri jt väiteid, et asümmeetriliste mõõdikute kasutamine on vajalik just seetõttu, et need

tuvastavad asümmeetrilisi seoseid paremini kui sümmeetrilised mõõdikud ning annavad uurijatele rohkem psühholoogiliselt relevantset informatsiooni uuritavate sõnaühendite kohta, sest inimese meeles on seosed nii sümmeetrilised kui ka asümmeetrilised.

Kokkuvõte

Magistritöö käsitles sõnadevahelise seose tugevuse mõõtmise meetodeid eesti keele ühendverbide automaatsel tuvastamisel tekstikorpusest. Eesmärk oli välja selgitada parim sõnadevahelise seose tugevuse mõõtmise meetod eesti keele ühendverbide automaatseks tuvastamiseks keelekorpusest ning, teada saada, kas mõõdikute tööd mõjutab tekstiliik ja/või korpuse suurus. Samuti oli eesmärk testida asümmeetrilisi statistikuid ja uurida, kas ja kui otstarbekas on nende rakendamine eesti keele ühendverbide tuvastamisel tekstikorpusest.

Töö esimene osa andis ülevaate uurimuse teoreetilistest ja praktilistest eeldustest. Täpsemalt kirjeldasin püsiühendi, kollokatsiooni ja ühendverbi mõisteid, lähenemisviise sõnadevahelise seose mõõtmiseks ja seda, missugust andmestikku sõnadevahelise seose tugevuse mõõtmise meetodite rakendamiseks vaja on. Lõpetuseks andsin põgusa ülevaate varasematest eesti keele püsiühendite automaatse tuvastamise katsetest. Töö teine osa esitas ülevaate sõnadevahelise seose tugevuse mõõtmisest: sellest, milline oli selles töös kirjeldatud katse materjal, kuidas seda koguti ja töödeldi, millised olid tulemuste hindamise meetodid. Samuti kirjeldasin võrdlusesse kaasatud sümmeetrilisi ja asümmeetrilisi mõõdikuid. Kolmandas peatükis esitasin mõõdikute tulemuste võrdlused sõltuvalt tekstiklassist ning korpuse suurusest. Lõpetuseks võrdlesin sümmeetriliste ja asümmeetriliste statistikute tulemusi ning kõrvutasin selle töö tulemusi sarnaste muude keelte andmestike peal tehtud katsete tulemustega.

Töö tulemused lükkavad ümber hüpoteesi, et parimaks meetodiks eesti keele ühendverbide tuvastamisel on log-tõepära funktsioon. Kõikide katsete kokkuvõttena saab väita, et kõige edukam statistik on t-skoor. Siiski töös esitatud erinevad võrdlused kinnitavad fakti, et ühtegi mõõdikut ei saa pidada teistest ühemõtteliselt paremaks, sest tulemusi mõjutab nii see, millist tekstiliiki sisaldab ühendverbide tuvastamiseks kasutatav korpus ning kui suur see korpus on, aga ka vaadeldavate kandidaatpaaride hulk.

Selgus, et korpuse tekstiliik mõjutab statistikute tulemusi ning oluline on arvestada sellega, et tulemused muutuvad kandidaatpaaride arvu muutudes. 50 kandidaatpaari seas

on t-skoor kõikide tekstiklasside korral parim statistik. 2000 kandidaatpaari hulgas on t-skoor parim mõõdik ühendverbide tuvastamisel küll aja- ja teaduskirjandustekstidest, kuid mitte ilukirjandustekstidest, millest sai ühendverbide tuvastamisega kõige paremini hakkama lihtne sagedusloend. Tulemuslikud on ka log-tõepära funktsioon, MS ja hii-ruut-statistik ning asümmeetrilised mõõdikud: tinglik tõenäosus ja ΔP. Selgus, et vaadeldavate kandidaatpaaride arvu kasvamisel muutuvad statistikute tulemused erinevalt: kui teiste mõõdikute täpsused langevad kandidaatpaaride arvu suurenemisega, siis MI täpsus paraneb.

Ka korpuse suurusel on mõju mõõdikute tulemustele ühendverbide tuvastamisel: t-skoori, log-tõepära funktsiooni, hii-ruut-statistiku, MS-i, tingliku tõenäosuse ja ΔP tulemused paranevad korpuse suuruse kasvamisega. Vastupidise efekti annab korpuse mahu kasvamine MI tulemustele. Korpuse mahu suurenemine ei mõjuta lihtsa sagedusloendi tulemusi. Siiski kõige efektiivsemaks sümmeetriliseks statistikuks võib pidada t-skoori, millele järgneb log-tõepära funktsioon. Meetodite paremusjärjestust muudab erineva arvu kandidaatpaaride vaatlemine: kui 100 kandidaatpaari seas on paremuselt kolmas statistik MS, siis 2000 kandidaatpaari hulgas on selleks lihtne sagedusloend. Olenemata korpuse suurusest on ühendverbide tuvastamisel tulemuslik ka hii-ruut-statistik, kõige madalama täpsuse ja saagisega tuvastab ühendverbe erineva suurusega korpustest MI. Kui 5 miljoni sõna suuruses korpuses on 2000 kandidaatpaari seas lihtne sagedusloend log-tõepära funktsioonist parem, siis korpuse suurenemisega log-tõepära funktsiooni tulemused paranevad ning alates 10 miljoni sõna suurusest korpusest on log-tõepära funktsiooni tulemused paremad kui lihtsa sagedusloendi omad. Asümmeetrilised ΔP ja tinglik tõenäosus on samuti tulemuslikud ülesande lahendamisel igas suuruses korpusest.

Asümmeetriliste statistikute rakendamine ja võrdlemine sümmeetriliste statistikute tulemustega kinnitas eeldust, et asümmeetriliste mõõdikute rakendamine on tulemuslik ühendverbide tuvastamisel tekstikorpusest, sest need tuvastavad arvestatava arvu õigeid ühendverbe ning lisaks sisaldavad informatsiooni ühendisse kuuluvate sõnade seose suuna kohta. Kuigi mõõdikute võrdluses andis parima tulemuse sümmeetriline statistik t-skoor, siis töös kirjeldatud katsed tõestasid, et parima lõpptulemuse saamiseks on mõistlik kombineerida sümmeetrilisi ja psühholoogiliselt tugevama alusega asümmeetrilisi

statistikuid, sest sümmeetrilised mõõdikud ei tuvasta sõnadevahelisi asümmeetrilisi seoseid, mis on iseloomulikud inimese kognitiivsetele võimetele.

See töö täidab eesti keele korpuslingvistiliste uurimuste seas lünga, mis puudutab sõnadevahelise seose tugevuse mõõtmise statistilisi meetodeid laiemalt kollokatsioonide, kitsamalt ühendverbide automaatsel tuvastamisel. Sõnadevahelise seose tugevuse mõõtmine pole eesti keele kollokatsioonide tuvastamisel küll uus valdkond, kuid selleteemaline süstemaatiline uurimus siiani puudus. Kuigi selle töö eesmärgiks polnud hinnata, kas tuvastatud ühendit pidada ühendverbiks või mitte, võivad töö käigus loodud mõõdikute pingeread olla toetuspunktiks leksikograafidele, et otsustada, kas sõnaühend on ühendverb ja kas see peaks sõnaraamatusse kuuluma või mitte.

Sõnadevahelise seose tugevuse mõõtmise sümmeetrilisi statistikuid on teiste keelte peal testitud rohkelt ning üldine metodoloogia on siinsesse töösse üle kantud. Pisut uuenduslikum on asümmeetriliste statistikute rakendamine kollokatsioonide tuvastamisel, mida muude keelte peal väga palju katsetatud pole. Seetõttu on peatükis 2.1.4 esitatud sümmeetriliste ja asümmeetriliste statistikute võrdlemise meetodid küll selle töö eesmärkide täitmiseks piisavad, kuid vajavad kindlasti edasi arendamist.

Tulevikus on otstarbekas testida erinevate meetodite kombineerimise võimalusi ja hinnata nende kombinatsioonide tulemuslikkust kollokatsioonide tuvastamisel tekstikorpusest. Samuti vajavad põhjalikumat uurimist asümmeetriliste statistikute kasutamisvõimalused korpuse materjali peal, et täiendada psühholingvistiliste katsete tulemusi korpuslingvistiliste uurimuste tulemustega.

Kirjandus

Aavik, Johannes 1974. Keeleuuenduse äärmised võimalused. Eesti Keele ja Kirjanduse Instituudi toimetised 15. Stockholm: Eesti Keele ja Kirjanduse Instituut.

Bell, Alan, Jason M Brenier, Michelle Gregory, Cynthia Girand, Dan Jurafsky 2009. Predictability effects on durations of content and function words in conversational English. Journal of Memory and Language 60(1). 92–111.

Church, Kenneth Ward, Patrick Hanks 1990. Word association norms, mutual information, and lexicography. Computational linguistics 16(1). 22–29.

Daille, Béatrice 1996. Study and implementation of combined techniques for automatic extraction of terminology. The balancing act: Combining symbolic and statistical approaches to language 1. 49–66.

Dunning, Ted 1993. Accurate methods for the statistics of surprise and coincidence. Computational linguistics 19(1). 61–74.

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

EKK = *Erelt, Mati, Tiiu Erelt, Kristiina Ross* 2007. Eesti keele käsiraamat. Eesti Keele Sihtasutus. http://www.eki.ee/books/ekk09/.

EKSS = Eesti keele seletav sõnaraamat. http://www.eki.ee/dict/ekss/.

Ellis, Nick C 2006. Language acquisition as rational contingency learning. Applied Linguistics 27(1). 1–24.

Ellis, Nick C, Fernando Ferreira-Junior 2009. Constructions and their acquisition Islands and the distinctiveness of their occupancy. Annual Review of Cognitive Linguistics 7(1). 187–220.

Erelt, Mati 2013. Eesti keele lauseõpetus: Sissejuhatus. Öeldis. (Tartu Ülikooli Eesti Keele Osakonna Preprindid). Tartu: Tartu Ülikool.

Evert, Stefan 2004. The statistics of word cooccurrences. Dissertation. Stuttgart: Stuttgart University.

Evert, Stefan 2008. Corpora and collocations. Corpus Linguistics. An International Handbook 2. 223–233.

Evert, Stefan, Brigitte Krenn 2001. Methods for the qualitative evaluation of lexical association measures. Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, 188–195. Association for Computational Linguistics.

Garside, Roger, Geoffrey N Leech, Tony McEnery 1997. Corpus annotation: linguistic information from computer text corpora. Longman London, New York.

Gries, Stefan Th 2012. Corpus linguistics, theoretical linguistics, and cognitive/psycholinguistics: towards more and more fruitful exchanges. Language and Computers 75(1). 41–63.

Gries, Stefan Th 2013. 50-something years of work on collocations: What is or should be next. International Journal of Corpus Linguistics 18(1). 137–166.

Kaalep, Heiki-Jaan 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. Keel ja Kirjandus 1(1998). 22–29.

Kaalep, Heiki-Jaan, Kadri Muischnek 2002. Püsiühendite leidmine teksti abil. Tähendusepüüdja/Catcher of the Meaning, TÜ üldkeeleteaduse õppetooli toimetised 3. 172–184.

Kaalep, Heiki-Jaan, Kadri Muischnek 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. Eesti Rakenduslingvistika Ühingu aastaraamat 5. 157–172.

Kaalep, Heiki-Jaan, Kadri Muischnek 2012. Osalausete tuvastamine eestikeelses tekstis kui iseseisev ülesanne. Eesti Rakenduslingvistika Ühingu aastaraamat(8). 55–68.

Kaalep, Heiki-Jaan, Tarmo Vaino 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. Keel ja Kirjandus 1. 30–38.

Kis, Balázs, Begoña Villada, Gosse Bouma, Gábor Ugray, Tamás Bíró, Gábor Pohl, John Nerbonne 2003. Methods for the Extraction of Hungarian Multi-Word Lexemes. CLIN.

Krenn, Brigitte 2000. Empirical implications on lexical association measures. Proceedings of The Ninth EURALEX International Congress.

Krenn, Brigitte, Stefan Evert 2001. Can we do better than frequency? A case study on extracting PP-verb collocations. Proceedings of the ACL Workshop on Collocations, 39–46.

Manning, Christopher D, Hinrich Schütze 1999. Foundations of statistical natural language processing. MIT press.

Michelbacher, Lukas, Stefan Evert, Hinrich Schütze 2007. Asymmetric association measures. Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007).

Michelbacher, Lukas, Stefan Evert, Hinrich Schütze 2011. Asymmetry in corpus-derived and human word associations. Corpus Linguistics and Linguistic Theory 7(2). 245–276.

Muischnek, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles. Tartu: Tartu Ülikooli kirjastus.

Muuk, Elmar 1938. Verbide ja verbaalnoomenite kokkukirjutamisest. Eesti Keel 1. 4–17.

Pecina, Pavel, Pavel Schlesinger 2006. Combining association measures for collocation extraction. Proceedings of the COLING/ACL on Main conference poster sessions, 651–658.

Pedersen, Ted 1998. Dependent bigram identification. AAAI/IAAI, 1197.

Pedersen, Ted, Rebecca Bruce 1996. What to infer from a description. Technical Report 96-CSE-04. Southern Methodist University. Dallas, TX.

Pihlak, Ants 1985. Eesti ühendverbid ja perifrastilised verbid aspektitähenduse väljendajana. Ars Grammatica. Toim. Mati Erelt ja Henno Rajandi. Tallinn: Valgus. 62–93.

R Development CoreTeam 2011. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Ramisch, Carlos, Paulo Schreiner, Marco Idiart, Aline Villavicencio 2008. An evaluation of methods for the extraction of multiword expressions. Proceedings of the LREC Workshop-Towards a Shared Task for Multiword Expressions (MWE 2008), 50–53.

Rätsep, Huno 1969. Ühendverbide rektsioonistruktuuri iseärasusest eesti keeles. Emakeele Seltsi aastaraamat 14-15. 59–77.

Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. Valgus.

Sag, Ivan A, Timothy Baldwin, Francis Bond, Ann Copestake, Dan Flickinger 2002. Multiword expressions: A pain in the neck for NLP. Computational Linguistics and Intelligent Text Processing, 1–15. Springer.

Seretan, Violeta 2011. Syntax-based Collocation Extraction. (Text, Speech and Language Technology 44). Switzerland: Springer.

Sinclair, John 1991. Corpus, concordance, collocation. Oxford: Oxford University Press.

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

Valgma, Johannes, Nikolai Remmel 1968. Eesti keele grammatika. Tallinn: Valgus.

Wermter, Joachim, Udo Hahn 2006. You can't beat frequency (unless you use linguistic knowledge): a qualitative evaluation of association measures for collocation and term extraction. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, 785–792. Association for Computational Linguistics.

Wiechmann, Daniel 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. Corpus linguistics and linguistic theory 4(2). 253–290.

Statistical methods for particle verb extraction. *Summary*

The present master's thesis compares lexical association measures (AMs) for automatic extraction of Estonian particle verbs from the text corpus.

The main purpose of this study is to ascertain the best AM for Estonian particle verb extraction. In order to achieve the aim two comparisons have been submitted. The first one compares the performance of the AMs according to the text type in the corpus. This comparison is based on the Balanced Corpus of Estonian (15 million word tokens) containing newspaper, fiction and scientific texts. The second comparison, which focuses on the impact of the corpus size on the performance of the AMs, is based on the newspaper part (170 million words) of Estonian Reference Corpus. In addition, asymmetrical AMs have been involved into the study to observe their suitability for Estonian particle verb extraction.

The first chapter of the thesis contains theoretical overview of the extraction of multi-word expressions. The second chapter gives the review of the data used in the current work and introduces the symmetrical and asymmetrical association measures being evaluated. The third chapter is the practical part of the study. It contains the above-mentioned comparisons. In addition, the comparison of the performance of symmetrical and asymmetrical AMs is given. Finally, the results of current work has been compared to the results of previous studies on other languages.

The analysis of the first comparison reveals that the text type has an impact on the performance of the AMs. For 50 candidate pairs the best AM for Estonian particle verb extraction is t-score, which achieves the highest precision and recall in newspaper corpora. The performance of log-likelihood is the second best and it is followed by the MS, chi-squared statistic, (simple) frequency and MI. For 2000 candidate pairs t-score is the best symmetric AM for particle verb extraction from newspaper and scientific text corpus. t-score is followed by simple frequency which is the best AM for extracting particle verbs from fiction text corpus. Log-likelihood, MS and chi-squared statistic are

also suitable for extracting particle verbs from text corpus irrespective of the text type. The results of MI are lower than others. MI is suitable for extracting particle verbs from corpus when the number of candidate pairs is higher than 4000.

The results of the asymmetrical AMs also differ according to the text type. Both, conditional probability and ΔP, extracted most true particle verbs from newspaper corpora and least from scientific texts corpus. However, the CP(verb|adverb) and ΔP(verb|adverb) extracted larger number of true particle verbs than CP(adverb|verb) and ΔP(adverb|verb). This is the result of the fact that CP(adverb|verb) and ΔP(adverb|verb) raise rare word pairs that contain infrequent or even grammatically incorrect verbs. Thus, CP(adverb|verb) and ΔP(adverb|verb) are suitable for extracting rare word pairs and less-common particle verbs. All in all, symmetrical conditional probability and ΔP are suitable for the task of Estonian particle verb extraction from text corpora containing different text types.

The analysis of the second comparison reveals that the corpus size has an impact on the performance of the AMs. For the 100 candidate pairs the precision of t-test is the highest and it does not change with the respect to the corpus size. The log-likelihood is the second best and it is followed by the MS. For 5 and 20 million word corpus the (simple) frequency is better than chi-square, but for 170 million word corpus the chi-square is better than frequency. The MI has the worst results. For the 2000 candidate pairs the results are different than for n=100. Though the results of t-test are the best, irrespective of the corpus size and the performance of log-likelihood is the second best as the corpus size increases. The results of (simple) frequency are similar to the t-test and log-likelihood and it is better than the MS, chi-squared statistic and MI for the smallest dataset as well as for the biggest dataset. The MS as a whole produces better results than chi-squared statistic, but as expected, the precision of MI is significantly lower than others.

For the 100 and 2000 candidate pairs the performance of (simple) frequency do not change significantly as the size of the corpus increases. The performance of t-score, log-likelihood function, MS and chi-squared statistic conditional probability and ΔP increase as the size of the corpus increases and the precision of MI decreases as the corpus size

increases. The expansion of the corpus size affects least the performance of the (simple) frequency. All in all, corpus size has an impact on the performance of AMs.

The comparison of symmetrical and asymmetrical AMs revealed that asymmetrical association measures are suitable for the task of Estonian particle verb extraction and provide us slightly different and more detailed information about the extracted particle verbs. The results presented in this thesis prove that further study of asymmetrical AMs is necessary and more experiments are needed to broaden the knowledge about the performance of asymmetrical AMs.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina ELERI AEDMAA

(*autori nimi*)

(sünnikuupäev: 24.05.1989)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

„Sõnadevahelise seose tugevuse mõõtmise statistilised meetodid ühendverbide tuvastamisel”,

(*lõputöö pealkiri*)

mille juhendajad on Kadri Muischnek ja Kristel Uiboaed,

(*juhendajate nimed*)

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, _____ (22.05.2014)