doi:10.5128/ERYa11.03

STATISTILISED MEETODID ÜHENDVERBIDE TUVASTAMISEL TEKSTIKORPUSEST

Eleri Aedmaa

Ülevaade. Artiklis võrdlen sõnadevahelise seose tugevuse mõõtmise statistilisi meetodeid, mida kasutatakse arvutilingvistikas püsiühendite tuvastamiseks. Töö põhieesmärk on rakendada viit sümmeetrilist statistikut – t-skoori, vastastikuse informatsiooni väärtust, hii-ruut-statistikut, log-tõepära funktsiooni ja minimaalset tundlikkust – erineva suurusega korpuste peal ja välja selgitada, milline meetod töötab eesti keele ühendverbide automaatsel tuvastamisel kõige paremini. Teine suurem eesmärk on katsetulemuste põhjal uurida, milline on korpuse suuruse mõju statistikute tööle. Lisaks palju testitud nimetatud sümmeetrilistele statistikutele rakendan psühholoogiliselt paremini põhjendatud asümmeetrilisi statistikuid – tinglikku tõenäosust ja ΔP-d – ning toon välja nende eelised sümmeetriliste statistikute ees.*

Võtmesõnad: arvutilingvistika, korpuslingvistika, püsiühendid, statistika, ühendverbid, eesti keel

1. Sissejuhatus

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 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), mida näiteks Kaalep ja Muischnek (2002: 173) defineerivad kui sõnaühendit, millesse kuuluvad sõnad esinevad tekstis koos sagedamini, kui võiks eeldada nende eraldiesinemise sageduse põhjal. Sinclairi (1991: 71) definitsiooni järgi on kollokatsioon kombinatsioon kahest sõnast, mis näitavad tendentsi esineda koos loomuliku keele tekstides.

Keele automaattöötlemisel on kollokatsioonid problemaatiline nähtus, sest süntaktilise ja semantilise analüüsi jaoks on oluline mitmesõnalise üksuse või

* Artikkel põhineb autori magistritööl (Aedmaa 2014a).

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 statistikuid ehk sõnadevahelise seose tugevuse mõõdikuid (ingl *association measures*), mis on statistilised valemid sõnadevahelise seose statistilise tugevuse arvutamiseks (Evert 2008: 5). Artiklis kasutatakse mõisteid *statistik* ja *mõõdik* sünonüümselt. Statistikud on osutunud tulemuslikuks arvukates kontekstides: näiteks lähisünonüümide eristamisel leksikograafias või leksikaal-semantilistes uurimustes, andmekaeves ja masintõlkega seotud ülesannetes (Wiechmann 2008: 257).

Viimase 50 aasta jooksul on enim testitud sümmeetrilisi statistikud (Gries 2013: 4), mis arvutavad igale korpusest leitud sõnapaarile ühe seose tugevuse väärtuse (ingl *association score*), mis näitab kahe sõna vahelise statistilise seose suurust (Evert 2008: 5). Lisaks sümmeetrilistele mõõdikutele on viimastel aastatel kollokatsioonide tuvastamisel arvestatud ka sõnadevahelise seose asümmeetrilisusega, mille tuvastamiseks rakendatakse asümmeetrilisi sõnadevahelise seose tugevuse mõõdikuid. Nende abil on võimalik arvutada igale sõnapaarile kaks seose tugevuse väärtust, mis osutavad, milline sõna kollokaadis on rohkem mõjutatud teise sõna esinemisest ehk näitavad kollokatiivsuse suunda. (Gries 2013: 5–13)

Sümmeetrilisi sõnadevahelise seose tugevuse mõõdikuid on eesti keele ühendverbide peal varem rakendanud Uiboaed (2010), kes tuvastas Eesti murrete korpuse kaheliikmelisi ühendverbe ning katsetas kolme murderühma peal eraldi nelja statistikut. Uurimus kinnitas, 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 projekti "Eesti keele koondkorpuse esituse ja kasutusvõimaluste arendamine" raames automaatne veebis kasutatav kollokatsioonide leidja Tasakaalus korpusest[1], kus on rakendatud kolme statistikut. Sümmeetrilisi mõõdikuid on rakendatud ka teistsuguste ülesannete lahendamiseks, näiteks Jelena Kallas (2013) katsetas neid eesti keele sisusõnade süntagmaatiliste suhete tuvastamiseks.

Oma töös rakendan nii sümmeetrilisi kui ka asümmeetrilisi mõõdikuid, et tuvastada püsiühendite hulka kuuluvaid kahest komponendist – afiksaaladverbist ja verbist – koosnevaid ühendverbe. Artikkel on üles ehitatud järgnevalt. Esmalt teen ülevaate sõnadevahelisest seosest ja selle tugevuse mõõtmisest, millele järgneb materjali ja selle töötlemist ning tulemuste hindamist kirjeldav peatükk. Seejärel tutvustan rakendatud meetodeid ning viimaks analüüsin ja võrdlen sümmeetriliste ja asümmeetriliste statistikute töö tulemusi.

2. Sõnadevaheline seos ja selle tugevuse mõõtmine

Ainult sõnade kordumine ei ole piisav alus, et sõnadevahelist seost tugevaks pidada. Sõnadevahelise seose tugevuse statistikute rakendamine on vajalik, sest need aitavad määrata, kas tegemist on n-ö õige püsiühendiga ja kas sõnade vahel on tugev või nõrk seos. (Evert 2008: 5) Koosesinevate sõnade vahelise seose tuvastamiseks

[1] https://korpused.keeleressursid.ee/clc/ (16.12.2014).

saab rakendada matemaatilisi seose tugevuse mõõdikuid, mis annavad iga sõnapaari jaoks seose tugevuse skoori. Seose tugevuse skoori abil saab välja valida n-ö õiged püsiühendid rakendades sageduse lävendit või järjestades sõnapaarid seose tugevuse väärtuse järgi kahanevalt, mille tulemusena leiab n-ö õiged püsiühendid nimekirja eesotsast. (Pecina, Schlesinger 2006: 652)

Selleks et sõnadevahelise seose tugevuse mõõtmise statistikuid rakendada, tuleb defineerida, mida tähendab sõnade koosesinemine. Ka peab otsustama, 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 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: 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, kuid see pole 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 ei ole püsiühend. 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 on sõnadevahelise seose tugevuse mõõtmiseks kasutatavate statistikute jaoks vajalik andmestik järgmine: O on sõnade koosesinemise sagedus valimis, f_1 ja f_2 on vastavalt sõnapaari kuuluva esimese ja teise sõna marginaalsagedus ja N on valimimaht (Evert 2004: 36). Lisaks nõuavad statistilised meetodid ka teoreetilist sagedust (ingl *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 (ingl *contingency table*), mis arvestab marginaalsagedustega. Tabel 1 esitab näite kahemõõtmelisest sagedustabelist koos tehetega, mis illustreerivad vajalike väärtuste leidmist.

Tabel 1. Kahemõõtmeline sagedustabel (Evert 2008)

O11	**O**	**f1 – O**	O12
O21	**f2 – O**	**N – f1 – f2 + O**	O22

3. Materjal ja tulemuste hindamine

3.1. Materjal

Selle töö uurimismaterjal on Eesti keele koondkorpuse[2] ajakirjandustekstid – kokku 170 miljonit sõna. Materjal oli eelnevalt programmiga t3mesta (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, mille korral moodustavad kaks sõna potentsiaalse püsiühendi, kui nad esinevad ühes ja samas tekstiüksuses, tüüpiliselt lauses või lausungis (Evert 2008: 13), minu töö puhul osalauses, sest ühendverbi moodustavad sõnad saavad esineda vaid samas osalauses. Teksti sõnajärge kandidaatpaaride moodustamisel ma ei arvestanud ning kuna eesmärk oli tuvastada ühendverbe, siis iga osalause sees genereeriti adverbi ja verbi kõikvõimalikud kombinatsioonid, mis moodustavad kandidaatpaaride loetelu. Selle nimekirja peal rakendasin stopp-sõnade loendit ehk eemaldasin ühendid, milles esinev adverb reeglina ühendverbi koosseisus ei esine (nt *ikka*, *jälle*). Stopp-sõnade loend põhineb "Eesti keele seletava sõnaraamatu" (EKSS) ühendverbide loendil[3], mis ühtlasi on selles töös ka kuldstandardiks ehk õigete ühendverbide loeteluks, millega saadud tulemusi võrdlesin. EKSS-i ühendverbide loendist eemaldasin ühendverbid, mis ei ole selles töös kasutatud korpuse märgendamise seisukohalt võimalikud. Sellisteks ühenditeks on sõnavormiga *kätte*, *minema* ja *tulema* moodustatud ühendverbid, nt *kätte jõudma*, *kätte maksma*, *minema kihutama*/*minema*/*viskama* ja *tulema tulema*. *Kätte, minema* ja *tulema* 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 ühendit ning nii sisaldab selles töös kasutatud 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.

3.2. 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.

[2] Eesti keele koondkorpus. http://www.cl.ut.ee/korpused/segakorpus/index.php (16.12.2014).

[3] EKSS-i ühendverbide loetelu ja artiklis esitatud tulemuste põhjalikumad ülevaated on kättesaadavad aadressil kodu.ut.ee/~eleriaed/magistrit88_failid/ (16.12.2014).

Tabel 2 esitab siinse töö materjali andmed: materjal koosneb 170 miljonist sõnast ja 25 912 251 osalausest, materjalist genereeriti 67 558 kandidaatpaari ning õigeid ühendverbe on nende seas 1676. Kokku oli õigeid ühendverbe võimalik tuvastada 1737.

Tabel 2. Ülevaade töös kasutatud materjalist

sõnu	170 000 000
osalauseid	24 322 394
kandidaatpaare	67 558
tuvastatud õigeid ühendverbe	1676
õigeid ühendverbe kuldstandardis	1737
täpsus	(1676/67558)*100 ≈ 2,5%
saagis	(1676/1737)*100 ≈ 96,5%

Sageduse täpsus kogu materjalist ühendverbide tuvastamisel näitab, et 2,5% leitud ühenditest on EKSS-i ühendverbid. Saagise väärtus väljendab, et tuvastati 96,5% kõigist EKSS-i ühendverbidest ehk tuvastamata jäi 61 EKSS-i nimistusse kuuluvat ühendverbi. Sagedusloendi kasutamine tagab suure hulga õigete ühendverbide leidmise, kuid tuvastatud ühendite koguhulgast moodustavad õiged ühendverbid väikese osa.

3.3. Täpsuse kõverad

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

Näiteks võetakse 500 esimest sõnapaari, mis on saanud t-skoor statistiku suuremad 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 mõõta uuritavate meetodite tulemuslikkust püstitatud ülesande lahendamisel, arvutatakse välja algtaseme täpsus (ingl *baseline precision*) ehk teoreetiline täpsus, mis näitab, kui suur osa kõikidest leitud ühenditest on õiged ühendverbid. Algtaseme täpsus arvutatakse korpusest leitud õigete ühendverbide arvu jagamisel genereeritud kandidaatpaaride arvuga. Seejärel kantakse saadud tulemused täpsuse kõveratena joonisele, kus iga kõver märgib erineva 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 (Krenn, Evert 2001, Wermter, Hahn 2006) on tõestanud ka koosesinemise sageduse tõhusust püsiühendite tuvastamise katsetes, siis on siinses töös uuritud, millised on koosesinemise sageduse tulemused võrreldes teiste meetoditega.

4. Sõnadevahelise seose tugevuse mõõtmise meetodid

Sõnadevahelise seose tugevuse mõõdikuid on palju ja nende rakendusalad lingvistikas erinevad, kuid kõige rohkem on neid kasutatud kollokatiivsust käsitlevates uurimustes (Wiechmann 2008: 254–257). Eelnevad tööd on 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).

4.1. Sümmeetrilised mõõdikud

Sümmeetrilised sõnadevahelise seose tugevuse mõõtmise statistikud 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.

Selles töös olen sümmeetrilistest statistikutest uurimiseks välja valinud kõige sagedamini sarnaste ülesannete lahendamiseks kasutatavad mõõdikud.

1. **t-skoor** (ingl *t-score*) (Church, Hanks 1990), mille arvutamisel lahutatakse sagedusest teoreetiline sagedus ning see tulemus jagatakse koosesinemise sageduse ruutjuurega:

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

t-skoor on osutunud kasulikuks näiteks saksa keele prepositsioonifraasi ja verbiühendite tuvastamisel (Krenn, Evert 2001).

2. **Vastastikuse informatsiooni väärtuse** (ingl *mutual information*, MI) (Church, Hanks 1990) arvutamiseks jagatakse osalausete hulk, kus mõlemad sõnapaari liikmed esinevad, teoreetilise sagedusega, ning sellest võetakse omakorda logaritm alusel kaks:

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

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

3. **Hii-ruut-statistiku** (ingl *chi-squared measure*) (Manning, Schütze 1999) väärtus arvutatakse kahemõõtmelise sagedustabeli väärtuste abil:

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

Hii-ruut-statistik on kahepoolne mõõdik, mis tähendab, et suure positiivse väärtuse saavad nii tugeva kui ka nõrga seosega sõnaühendid ehk suure 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)

4. **Log-tõepära funktsioon** (ingl *log-likelihood measure*) (Dunning 1993) on kõige laialdasemat kasutust leidnud statistik arvutilingvistikas (Evert 2008: 31). Sarnaselt hii-ruut-statistikule arvutatakse ka log-tõepära funktsiooni väärtused kahemõõtmelise sagedustabeli väärtuste põhjal:

$$\text{log-tõepära} = 2 \sum_{ij} O_{ij} \log \frac{O_{ij}}{E_{ij}}$$

Ka log-tõepära funktsioon on kahepoolne mõõdik (Evert 2008: 22) ning vajab ühepoolseks teisendamist (vt eelmine punkt).

5. **Minimaalne tundlikkus** (ingl *minimum sensitivity*, MS) (Pedersen 1998) on lihtne ja efektiivne mõõdik bigrammide tuvastamiseks. MS arvutatakse sõna bigrammis esinemise sageduse ja sama sõna üldise esinemissageduse võrdlemisel valemi põhjal:

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

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)

4.2. Asümmeetrilised mõõdikud

Eelnevalt 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* esinemise järgi. Kuid vastupidi see ei toimi: kui osalauses esineb *minema*, siis *lahku* seal sama suure tõenäosusega ei esine ehk seos *minema* ja *lahku* vahel nii tugev ei ole ning *lahku* esinemine ei ole hästi ennustatav *minema* esinemise järgi. Erinevalt sümmeetrilistest statistikutest ei ühenda asümmeetrilised statistikud 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 ja ΔP-d.

1. **Tinglik tõenäosus** (ingl *conditional probability*) (Bell jt 2009, Michelbacher jt 2007) on MS-ist tuletatud sõnadevahelise seose tugevuse mõõdik. Tinglik tõenäosus arvutatakse kahe valemi abil:

$$p(sõna_2|sõna_1) = \frac{O11}{O11 + O12}$$

$$p(sõna_1|sõna_2) = \frac{O11}{O11 + O21}$$

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

2. **ΔP** (Ellis 2006, Ellis, Ferreira-Junior 2009) 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:

$$\Delta P_{2|1} = p(sõna_2|sõna_1=esineb) - p(sõna_2|sõna_1 = puudub) = \frac{O11}{O11 + O12} - \frac{O21}{O21 + O22}$$

$$\Delta P_{1|2} = p(sõna_1|sõna_2 = esineb) - p\,(sõna_1|sõna_2 = puudub) = \frac{O11}{O11 + O21} - \frac{O12}{O12 + O22}$$

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 on leidnud kasutust psühholoogilistes uurimustes ja osutunud mõõdikuks, mis iseloomustab paremini inimese kognitiivseid võimeid.

5. Meetodite täpsused ja saagised erineva suurusega korpuste põhjal

Selleks et hinnata valimi suuruse mõju mõõdikute tulemustele, võtsin 170 miljoni sõna suurusest korpusest viis valimit: esimene koosnes 5 miljonist sõnast, teised vastavalt 10, 20, 70 ja 170 miljonist sõnast.

5.1. Sümmeetriliste statistikute ja koosesinemise sageduse tulemused

Tabel 3 esitab ülevaate koosesinemise sageduse tulemustest erineva suurusega korpustest ühendverbide tuvastamisel.

Tabel 3. Sageduse täpsus ja saagis erineva suurusega ajakirjanduskorpustes

Sõnu	Osalauseid	Kandidaatpaare	Õigeid ühendverbe	Täpsus	Saagis
5 mln	707 979	13 141	1351	10,3%	77,8%
10 mln	1 410 474	18 545	1459	7,9%	84,0%
20 mln	2 823 255	26 268	1532	5,8%	88,2%
70 mln	9 640 426	46 863	1628	3,5%	93,7%
170 mln	24 322 394	67 558	1676	2,5%	96,5%

Selgub, et näiteks 20 miljonist sõnast koosneva ajakirjandustekstide valimist 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. Mida suurem on korpus, seda rohkem kandidaatpaare genereeriti ja õigeid ühendverbe tuvastati. Seega on suurima korpuse puhul saagis kõige suurem (96,5%) ja täpsus kõige väiksem (2,5%).

Tabelis 4 on esitatud iga statistiku 25 suurima väärtuse saanud ühendit ning sealt on näha, et MI kõrgemad väärtused on saanud harva esinevad ühendid, millesse kuuluv verb esineb samuti harva. Seetõttu on MI hea harvaesinevate (õigete) ühendverbide tuvastamiseks. Koosesinemise sageduse eripärana saab välja tuua asjaolu, et kõrge väärtusega on *olema*-verbi sisaldavad ühendid, sest *olema* on korpuses kõige sagedasem verb. Ka selgub, et suured statistikute väärtused on saanud ühendid, mis EKSS-i ühendverbide loendisse ei kuulu, kuid millesse kuuluvate sõnade vaheline seos on tugev ehk neid võib pidada ühendverbideks, näiteks *kaasa võtma*, *kallale tungima*, *eemale peletama*, *edasi lükkuma*, *esile kerkima*. Kokku oli neid ühendeid umbes 30, kuid täpne arv sõltub sellest, millist seose tugevuse väärtust pidada tugevaks. Selliste ühendite esinemine kinnitab, et püsiühendite tuvastamine tekstikorpusest on kasulik leksikograafide töös.

Tabel 4. Statistikute 25 suurima väärtusega ühendit ajakirjandustekstides

t-skoor	MI	hii-ruut	log-tõepära	MS	sagedus
vastu võtma	*sekka jõratama*	*kallale tungima*	*vastu võtma*	*kallale tungima*	*välja olema*
ette nägema	*sekka mahakooruma*	*eemale peletama*	*ette nägema*	*ette nägema*	*üle olema*
välja tulema	*sekka taas-taas-taasavaldama*	*ette kujutama*	*kaasa tooma*	*eemale peletama*	*kokku olema*
kaasa tooma	*ühte kanseldama*	*kokku leppima*	*kokku leppima*	*ümber lükkama*	*ära olema*
kinni pidama	*järel mitte-kunagi-armastama*	*vastu võtma*	*ette kujutama*	*kaasa aitama*	*vastu võtma*
välja kuulutama	*tagant turgima*	*ette nägema*	*välja kuulutama*	*maha müüma*	*välja tulema*
kokku leppima	*kallale õppi-matma*	*ette valmistama*	*alla kirjutama*	*kaasa tooma*	*ette nägema*
läbi viima	*kallale tromama*	*kaasa tooma*	*ette valmistama*	*läbi viima*	*ette olema*
alla kirjutama	*valla portesteerima*	*esile tõstma*	*läbi viima*	*ette kujutama*	*läbi olema*
välja andma	*sekka needistama*	*alt vedama*	*kaasa aitama*	*alla kirjutama*	*valmis olema*
ette kujutama	*sekka tšekkama*	*edasi lükkama*	*kinni pidama*	*vastu võtma*	*välja andma*
ette võtma	*pärale muganema*	*alla kirjutama*	*maha müüma*	*ette valmistama*	*kinni pidama*
ette valmistama	*alt delegeeruma*	*edasi lükkuma*	*edasi lükkama*	*üles kutsuma*	*kaasa tooma*
tagasi tulema	*alt tsirukleerima*	*välja kuulutama*	*välja tulema*	*esile tõstma*	*tagasi olema*
alla jääma	*üleval walitsema*	*kokku põrkama*	*vahele jääma*	*edasi lükkama*	*kokku saama*
üle andma	*taha praeguma*	*ümber lükkama*	*kokku puutuma*	*kokku leppima*	*välja kuulutama*
kaasa aitama	*tasa posisema*	*külge pookima*	*ilma jääma*	*esile kerkima*	*vastu olema*
maha müüma	*taga garaaxima*	*läbi viima*	*ette võtma*	*üles astuma*	*ette võtma*
välja töötama	*taga pusletama*	*kokku puutuma*	*alla jääma*	*üles ehitama*	*tagasi tulema*
kaasa võtma	*taga nõduma*	*kaasa aitama*	*kinni maksma*	*kinni hoidma*	*läbi viima*
ära kasutama	*alt näpsima*	*maha müüma*	*esile tõstma*	*maha laskma*	*ära tegema*
maha võtma	*eemale seminaritsema*	*kõrvale hiilima*	*tagasi lükkama*	*alt vedama*	*ligi olema*
kinni maksma	*eemale otsas/ muigama*	*ette heitma*	*ette heitma*	*alla kukkuma*	*kokku leppima*
ilma jääma	*eemale jää-minema*	*taga ajama*	*kokku põrkama*	*kaasa lööma*	*alla kirjutama*
välja tooma	*eemale hoidsima*	*taga kiusama*	*üles kutsuma*	*tagasi astuma*	*üles olema*

Ülevaate korpuse mahu suurenemise mõjust mõõdikute tulemustele ja paremusjärjestusele saab tabelist 5, kus on esitatud sümmeetriliste statistikute ja koosesinemise sageduse täpsused ja saagised ühendverbide tuvastamisel erineva suurusega korpustest 100, 1000 ja 2000 statistikute suurima väärtuse saanud sõnapaari seas.

Tabel 5. Sümmeetriliste statistikute ja koosesinemise sageduse täpsused ja saagised erineva suurusega ajakirjanduskorpustes

stat*	Hindamise-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%

* Lühendid: stat – statistik, t – t-skoor, hii – hii-ruut-statistik, log – log-tõepära funktsioon, sag – koosesinemise sagedus.

Tabelis 5 esitatud tulemused kinnitavad, et t-skooril on kõige paremad tulemused ühendverbide tuvastamisel erineva suurusega korpustest. t-skoori tulemused paranevad korpuse kasvades: 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. Seega, mida suurem on korpus, seda paremad on t-skoori tulemused.

MI tulemused korpuse 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 väärtus väheneb 16,2%-lt 6,4%-le. Seega on korpuse suuruse kasvul tugev negatiivne mõju MI tulemustele.

Hii-ruut-statistiku, log-tõepära funktsiooni ja MS-i tulemused korpuse mahu suurenedes paranevad. Kui 5 miljonist sõnast koosneva korpuse 100 kandidaatpaari hulgas on hii-ruut-statistiku 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 hii-ruut-statistiku saagis paraneb 45,6%-ni. Log-tõepära funktsiooni tulemused paranevad korpuse kasvades 100 kandidaatpaari seas vähem, 2000 kandidaatpaari hulgas rohkem. 5 miljonist sõnast ja 170 miljonist sõnast koosnevatest korpustest ühendverbide tuvastamisel on 100 kandidaatpaari lõikes log-tõepära funktsiooni täpsused ja saagised võrdsed, 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%. 5 ja 170 miljoni sõna suuruse korpuse

võrdlemisel näeb, et 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. Järelikult, mida suurem on korpus, seda paremad on nii hii-ruut-statistiku, log-tõepära funktsiooni kui ka MS-i tulemused.

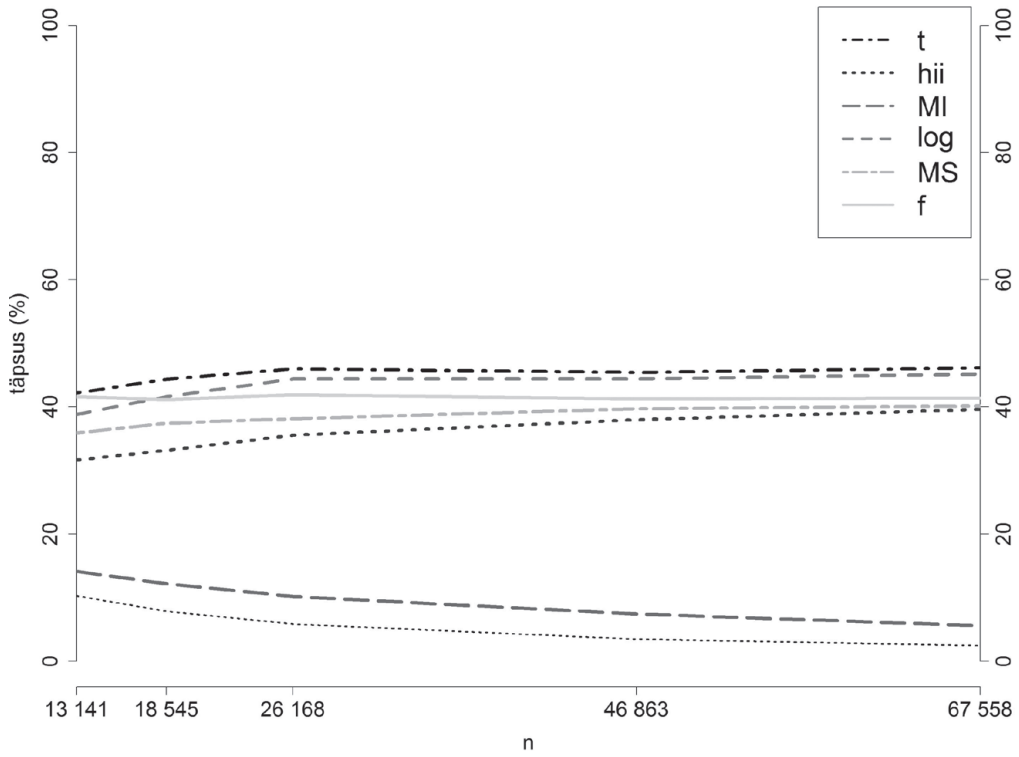
Koosesinemise sageduse tulemused ei muutu oluliselt: 100 kandidaatpaari seas on meetodi täpsus ja saagis korpuse suurusest olenemata sama. 2000 kandidaatpaari hulgas on täpsus ja saagis kõige suurem 20 miljonist sõnast koosneva valimi korral (vastavalt 41,9% ja 48,2%), 170 miljoni sõna suurusest korpusest ühendverbide tuvastamisel on koosesinemise sageduse täpsus 41,4%, mis erineb 5 miljoni sõna suurusele valimile vastavast täpsusest (41,6%) vaid 0,2%. Niisamuti on koosesinemise sageduse saagis 170 miljoni sõna suuruse korpuse korral (47,7%) võrreldes 5 miljoni sõna suuruse korpusega (47,9%) kahanenud 0,2%. Järelikult ei ole korpuse suurusel mõju koosesinemise sageduse tulemustele.

Korpuse suurusel ja ka vaadeldaval kandidaatpaaride arvul on mõju 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 koosesinemise sagedusest, mille tulemused on omakorda paremad kui hii-ruut-statistikul. 20 miljoni sõna suuruse korpuse puhul on koosesinemise sageduse ja hii-ruut-statistiku tulemused võrdsed ja 170 miljoni sõna suuruses korpuses on hii-ruut-statistiku tulemused koosesinemise sageduse omadest paremad. 2000 kandidaatpaari hulgas on üldjuhul log-tõepära efektiivsem kui koosesinemise sagedus, mille tulemused on omakorda paremad kui MS-i ja hii-ruut-statistiku omad. Vaid kõige väiksema korpuse korral on koosesinemise sagedus 2000 kandidaatpaari seas parem kui log-tõepära funktsioon.

Seda, kuidas mõõdikute tulemused muutuvad korpuse suurenedes, illustreerib joonis 1, mis esitab sümmeetriliste mõõdikute ja koosesinemise sageduse täpsuste muutused ühendverbide tuvastamisel erineva suurusega korpustest. Joonise x-telg märgib korpuse suurust korpusest genereeritud kandidaatpaaride arvu näol, y-telg märgib mõõdikute täpsust, kui arvesse võtta esimesed 2000 statistikute suurima väärtuse saanud kandidaatpaari. Peenike punktiirjoon joonisel tähistab algtaseme täpsust erineva suurusega korpustes.

Joonis 1 näitab, et kõige enam mõjutab 2000 kandidaatpaari hulgas korpuse kasv hii-ruut-statistiku tulemust, sest selle täpsus kasvab korpuse suurenemisega umbes 30%-lt 40%-ni. Samuti paranevad mõnevõrra t-skoori, log-tõepära funktsiooni ja MS-i tulemused. Koosesinemise sageduse tulemused korpuse mahu suurenemisega oluliselt ei muutu. MI tulemused halvenevad korpuse mahu suurendamisega.

Meetodite kõrvutamisel algtaseme täpsusega selgub, et kõikide mõõdikute täpsused on korpuse mahu kasvamisel suuremad kui algtaseme täpsus. Kui vaadelda esimest 2000 suurima statistikute väärtustega kandidaatpaari, siis üldjoontes on kõik mõõdikud eesti keele ühendverbide tuvastamisel tulemuslikud. Paremusjärjestus aga mõnevõrra muutub: kui 5 miljoni sõna suuruse korpuse korral on koosesinemise sagedus log-tõepära funktsioonist efektiivsem, siis suuremate korpuste korral on log-tõepära funktsiooni täpsus koosesinemise sageduse omast suurem. Ka võib öelda, et mida suurem on korpus, seda sarnasemad on koosesinemise sageduse, MS-i ja hii-ruut-statistiku tulemused. t-skoori täpsus on kõikide korpuste korral suurim, kuid suurema korpuse puhul on log-tõepära tulemused t-skoori tulemustega sarnasemad kui väiksema korpuse korral.



Joonis 1. Sümmeetriliste mõõdikute ja koosesinemise sageduse täpsused 2000 kandidaatpaari seas erineva suurusega ajakirjanduskorpustes

Tabeli 5 ja joonise 1 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. Koosesinemise sageduse tulemused jäävad korpuse kasvades peaaegu muutumatuks.

5.2. Asümmeetriliste mõõdikute tulemused

Selleks et uurida, 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 6, kus on esitatud asümmeetriliste mõõdikute tuvastatud õigete ühendverbide arvud erineva suurusega ajakirjandustekstide korpustest.

Tabel 6. Asümmeetriliste statistikute tuvastatud õigete ühendverbide hulk statistiku 50 suurima väärtuse saanud ühendi seas erineva suurusega korpustest

Sõnade arv korpuses	CP(verb|adverb)	CP(adverb|verb)	ΔP(verb|adverb)	ΔP(adverb|verb)
5 mln	25	1	36	1
10 mln	25	1	39	1
20 mln	26	0	38	0
70 mln	26	0	40	0
170 mln	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 suurima väärtuse saanud ühendi seas need numbrid vastavalt 28 ja 41. Vastupidist seost uurides aga tulemused korpuse mahu suurendes ei muutu või halvenevad. Seega oleneb korpuse mahu kasvu mõju asümmeetriliste statistikute puhul sellest, missugust seost sõnade vahel vaadelda. CP(adverb|verb) ja ΔP(adverb|verb) ei tuvasta palju õigeid ühendverbe, kuid sobivad harvaesinevate ühendite tuvastamiseks ja nad sisaldavad informatsiooni ühendite komponentide seoste kohta.

5.3. Sümmeetriliste ja asümmeetriliste mõõdikute tulemuste võrdlus

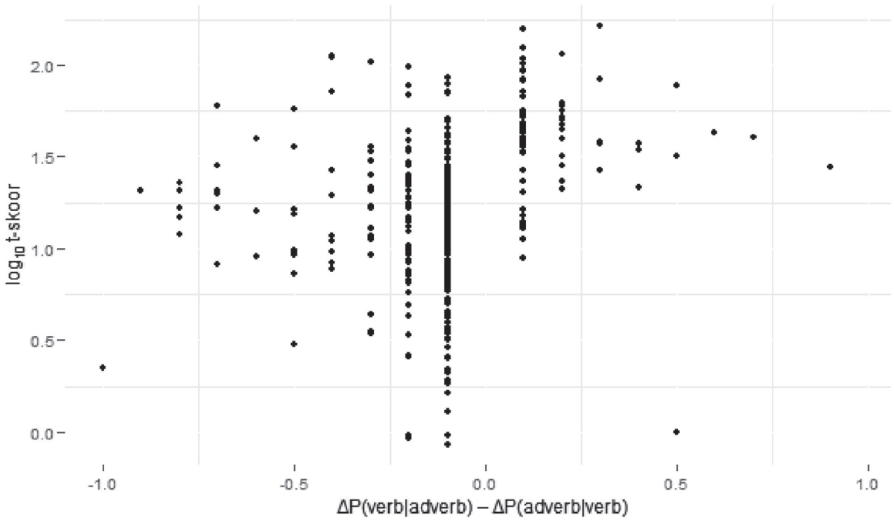
Sümmeetriliste ja asümmeetriliste mõõdikute tulemuste võrdluse eesmärgiks on välja selgitada, kas ja mida täpsemalt asümmeetriliste statistikute rakendamine lõpptulemusele juurde annab. Selleks võrdlen parimat sümmeetrilist statistikut – t-skoori – parima asümmeetrilise statistiku ΔP-ga (sümmeetriliste ja asümmeetriliste mõõdikute võrdlemisest saab põhjalikuma ülevaate autori artiklist Aedmaa 2014b).

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 t-skooriga arvestan mõlema statistiku 3314 suurima väärtuse saanud ühendit, mille seast eraldan õiged ühendverbid. ΔP hindamiseks võrdlen t-skoori tuvastatud õigete ühendverbide loendit ΔP tuvastatud õigete ühendverbide nimekirjaga.

t-skoori 3314 suurima 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.

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

Seda, et sümmeetrilised mõõdikud ei sisalda infot ühendite asümmeetrilisuse kohta, kinnitab joonis 2, 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 2. ΔP tuvastatud õigete ühendverbide jaotus ΔP(verb|adverb) – ΔP(adverb|verb) ja t-skoori väärtuste järgi

Jooniselt 2 selgub, et suure t-skoori väärtusega ühendeid leidub igasuguse ΔP väärtusega ühendite seas, mis tähendab, et t-skoor ja üldistatult ka teised sümmeetrilised statistikud omistavad ühenditele suuri kahesuunalisi väärtuseid arvestamata seda, kumb sõna teisest sõltub. Asümmeetrilise ΔP abil saab aga tuvastada ühendeid, mille komponentide vahel on asümmeetriline seos. Jooniselt 2 on näha ka see, 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 suurema 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. Samas teistest oluliselt paremat mõõdikut ei ole ja mõistlik on erinevat liiki statistikuid omavahel kombineerida.

6. Kokkuvõte

Artikkel käsitles sõnadevahelise seose tugevuse mõõtmise meetodeid eesti keele ühendverbide automaatsel tuvastamisel tekstikorpusest. Lisaks sümmeetrilistele statistikutele rakendasin ka kahte asümmeetrilist statistikut eesmärgiga uurida, kuidas erinevad sümmeetriliste ja asümmeetriliste statistikute tulemused.

Selgus, et t-skoori võib pidada kõige efektiivsemaks sümmeetriliseks statistikuks ühendverbide tuvastamisel ajakirjanduskorpusest, kuid kõiki rakendatud statistikuid tasub kasutada ühendverbide tuvastamiseks tekstikorpusest, sest korpuse suurus ja vaadeldavate kandidaatpaaride arv mõjutavad statistikute tööd.

Katsed tõestasid, et ka asümmeetriliste mõõdikute rakendamine on põhjendatud ühendverbide tuvastamisel tekstikorpusest, sest need tuvastavad arvestatava

arvu õigeid ühendverbe ning lisaks sisaldavad informatsiooni ühendisse kuuluvate sõnade seose suuna kohta. Parima lõpptulemuse saamiseks on mõistlik sümmeetrilisi ja asümmeetrilisi statistikuid kombineerida, sest sümmeetrilised mõõdikud ei tuvasta sõnadevahelisi asümmeetrilisi seoseid, mis on iseloomulikud inimese kognitiivsetele võimetele. Erinevate meetodite kombineerimise võimaluste ja nende kombinatsioonide tulemuslikkuse hindamine kollokatsioonide tuvastamisel tekstikorpusest jääb edasise uurimistöö teemaks.

Viidatud kirjandus

Aedmaa, Eleri 2014a. Sõnadevahelise seose tugevuse mõõtmise statistilised meetodid ühendverbide tuvastamisel. [Statistical Methods for Particle Verb Extraction.] Magistritöö. Käsikiri Tartu ülikooli üldkeeleteaduse osakonnas. http://hdl.handle.net/10062/44260

Aedmaa, Eleri 2014b. Statistical methods for Estonian particle verb extraction from text corpus. – Proceedings of the ESSLLI 2014 Workshop: Computational, Cognitive, and Linguistic Approaches to the Analysis of Complex Words and Collocations, 17–22.

Bell, Alan; Brenier, Jason M.; Gregory, Michelle; Girand, Cynthia; Jurafsky, Dan 2009. Predictability effects on durations of content and function words in conversational English. – Journal of Memory and Language, 60 (1), 92–111. http://dx.doi.org/10.1016/j.jml.2008.06.003

Church, Kenneth Ward; Hanks, Patrick 1990. Word association norms, mutual information, and lexicography. – Computational Linguistics, 16, 22–29.

Dunning, Ted 1993. Accurate methods for the statistics of surprise and coincidence. – Computational Linguistics, 19, 61–74.

Eesti keele seletav sõnaraamat. http://www.eki.ee/dict/ekss/ (16.12.2014).

Ellis, Nick C 2006. Language acquisition as rational contingency learning. – Applied Linguistics, 27 (1), 1–24. http://dx.doi.org/10.1093/applin/ami038

Ellis, Nick C.; Ferreira-Junior, Fernando 2009. Constructions and their acquisition: Islands and the distinctiveness of their occupancy. – Annual Review of Cognitive Linguistics, 7 (1), 188–221. http://dx.doi.org/10.1075/arcl.7.08ell

Evert, Stefan 2004. The Statistics of Word Cooccurrences. Dissertation. Stuttgart: Stuttgart University.

Evert, Stefan 2008. Corpora and collocations. – Anke Lüdeling, Merja Kytö (Eds.). Corpus Linguistics. An International Handbook 2. De Gruyter Mouton, 223–233. http://dx.doi.org/10.1515/9783110213881.2.1212

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. http://dx.doi.org/10.1075/ijcl.18.1.09gri

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

Kaalep, Heiki-Jaan; Muischnek, Kadri 2002. Püsiühendite leidmine teksti abil. [Extraction of multiword expressions using text corpus.] – Renate Pajusalu, Tiit Hennoste (Toim.). Tähendusepüüdja: pühendusteos professor Haldur Õimu 60. sünnipäevaks 22. jaanuaril 2002. Catcher of the Meaning: Festschrift for Professor Haldur Õim on the occasion of his 60th birthday. TÜ üldkeeleteaduse õppetooli toimetised 3. Tartu: Tartu Ülikool, 172–184.

Kaalep, Heiki-Jaan; Muischnek, Kadri 2009. Eesti keele püsiühendid arvutilingvistikas: miks ja kuidas. [Estonian multiword expressions in computational linguistics.] – Eesti

Rakenduslingvistika Ühingu aastaraamat, 5, 157–172. http://dx.doi.org/10.5128/ERYa5.10
Kaalep, Heiki-Jaan; Muischnek, Kadri 2012. Osalausete tuvastamine eestikeelses tekstis kui iseseisev ülesanne. [Clause splitting as a separate task (in the analysis of Estonian texts).] – Eesti Rakenduslingvistika Ühingu aastaraamat, 8, 55–68. http://dx.doi.org/10.5128/ERYa8.04
Kaalep, Heiki-Jaan; Vaino, Tarmo 1998. Kas vale meetodiga õiged tulemused? Statistikale tuginev eesti keele morfoloogiline ühestamine. [Getting correct results with an incorrect method? Morphological disambiguation of Estonian using statistics.] – Keel ja Kirjandus, 1, 30–38.
Kallas, Jelena 2013. Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias. [Syntagmatic Relationships of Estonian Content Words in Corpus and Pedagogical Lexicography.] Tallinna Ülikooli humanitaarteaduste dissertatsioonid 32. Tallinn: Tallinna Ülikool. http://e-ait.tlulib.ee/id/eprint/303
Krenn, Brigitte; Evert, Stefan 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.; Schütze, Hinrich 1999. Foundations of Statistical Natural Language Processing. Cambridge (Mass.)–London: MIT press.
Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich 2007. Asymmetric association measures. – Proceedings of the Recent Advances in Natural Language Processing (RANLP 2007).
Michelbacher, Lukas; Evert, Stefan; Schütze, Hinrich 2011. Asymmetry in corpus-derived and human word associations. – Corpus Linguistics and Linguistic Theory, 7 (2), 245–276. http://dx.doi.org/10.1515/cllt.2011.012
Pecina, Pavel; Sclesinger, Pavel 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; Bruce, Rebecca 1996. What to infer from a description. Technical Report 96-CSE-04. Southern Methodist University. Dallas, TX.
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. http://dx.doi.org/10.5128/ERYa6.19
Wermter, Joachim; Hahn, Udo 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.
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. http://dx.doi.org/10.1515/CLLT.2008.011

Eleri Aedmaa (Tartu Ülikool) uurimisvaldkonnad on korpuslingvistika, püsiühendid, statistilised meetodid keeleteaduses.
Tartu Ülikool, eesti ja üldkeeleteaduse instituut, Jakobi 2, 51014 Tartu, Estonia
eleriaed@ut.ee

STATISTICAL METHODS FOR ESTONIAN PARTICLE VERB EXTRACTION FROM TEXT CORPORA

Eleri Aedmaa
University of Tartu

The present article compares lexical association measures (AMs) for automatic extraction of Estonian particle verbs from the newspaper part of the Estonian Reference Corpus. The main purpose of this study is to ascertain the best symmetrical AM for Estonian particle verb extraction. The central focus lies on the impact of the corpus size on the performance of the compared symmetrical association measures. In addition, asymmetrical AMs have been included in the study to observe their suitability for Estonian particle verb extraction. Five symmetrical association measures have been used, namely the t-test, log-likelihood, χ^2, mutual information, and minimum sensitivity, as well as two asymmetrical association measures, namely conditional probability and ΔP. The association measures were compared against the co-occurrence frequency of verb and verbal particle.

The analysis of the comparison reveals that the t-test achieved the best precision values and the corpus size has an impact on the performance of the AMs. As the corpus size increased, the performances of the t-test, log-likelihood, χ^2 and minimum sensitivity increased and the precision of mutual information decreased. The performance of (simple) frequency did not change significantly as the size of the corpus increased.

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

Keywords: computational linguistics, corpus linguistics, multi-word expressions, particle verbs, statistics, Estonian