doi:10.5128/ERYa6.19

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 6, 307–326

STATISTILISED MEETODID MURDEKORPUSE ÜHENDVERBIDE TUVASTAMISEL

Kristel Uiboaed

Ülevaade. Sõnadevahelise seose tugevuse mõõtmise statistikuid kasutatakse arvutilingvistikas püsiühendite tuvastamisel. Statistikud võimaldavad korpuses kahele sõnale arvutada nendevahelise seose tugevuse väärtuse, mille põhjal võib otsustada, kas tegemist on püsiühendiga või mitte. Statistikute kasutamise eelis on, et arvesse ei võeta ainult sõnade koosesinemise, vaid ka ühendit moodustavate sõnade eraldiesinemise sagedusi. Artiklis teen katse rakendada statistikuid Eesti murrete korpuse kaheliikmeliste ühendverbide automaatsel tuvastamisel. Katsetatud on kolme murderühma peal eraldi nelja statistikut: t-skoori, vastastikuse informatsiooni väärtust MI, hii-ruut statistikut ning log-tõepära funktsiooni.*

Võtmesõnad: arvutilingvistika, korpuslingvistika, murdeuurimine, meetodid ja vahendid, statistika, eesti keel

1. Sissejuhatus

Keelekorpuste levik on teinud võimalikuks erinevate keelenähtuste uurimise suuremahulise materjali põhjal. See eeldab omakorda erinevate statistiliste meetodite kasutamist lingvistiliseks andmetöötluseks ning vajaduse korral uute rakenduste väljatöötamist. Oleks mõeldamatu, et inimene suudaks nii suurt andmestikku läbi vaadata ning seda analüüsida. Käesolev töö on esmane suurem katse rakendada arvutilingvistikas laialt levinud statistilisi meetodeid ka Eesti murrete korpuse[1] peal. Ühendverbide automaatseks tuvastamiseks Eesti murrete korpusest kasutatakse sõnadevahelise seose tugevuse mõõtmise statistikuid (ingl *association measures*) – statistilisi valemeid, mille abil on võimalik arvutada igale korpuses olevale sõnaühendile seose tugevuse väärtus (*association score*), mis osutab sõnadevahelise statistilise seose tugevusele. (Evert 2008)

* Artikli valmimist on toetanud riikliku programmi "Eesti keel ja kultuurimälu 2009" projekt EKKM09-111 ja riikliku programmi "Eesti keele keeletehnoloogiline tugi 2006 " projekt EKKTT06-14 ning Eesti Teadusfondi grant nr 7464. Tänan anonüümseid retsensente asjatundlike ettepanekute ja kommentaaride eest.

[1] Eesti murrete korpus www.murre.ut.ee (30.08.2009).

Erinevad statistikud põhinevaid erinevatel teooriatel, seega erinevate statistikute abil arvutatud sõnadevahelise seose tugevuse väärtused pole omavahel võrreldavad, kuna väärtuste leidmise alused pole ühtsed. Statistikud võivad põhineda lihtsalt sageduste kokkulugemisel, informatsiooniteoorial[2] või neil võib olla statistiline alus (Evert 2008, Krenn, Evert 2001).

Sõnadevahelise seose tugevuse mõõtmise statistikuid kasutatakse arvutileksikograafias, eelkõige püsiühendite tuvastamisel, ning paljudes arvutilingvistilistes rakendustes, näiteks masintõlkes. Statistikuid on võimalik rakendada muudegi grammatiliste konstruktsioonide ja lingvistiliste nähtuste uurimisel. Näiteks on statistikute abil võimalik uurida verbi ja objekti suhteid, verbide transitiivseid ja intransitiivseid omadusi (Stefanowitsch, Gries 2009).

Artikli eesmärk on tutvustada statistiliste meetodite tööpõhimõtteid ning rakendada neid murdekorpuse morfoloogiliselt märgendatud ja automaatselt osalausestatud tekstide peal. Esiteks tutvustan statistikute kasutamiseks vajalikke mõisteid ja andmeid, kirjeldan kolme võimalikku materjali liigendamisviisi ning näitan, kuidas arvulised väärtused valemitesse saadakse. Seejärel teen ülevaate kasutatud materjalist. Viimases osas kirjeldan nelja väljavalitud sõnadevahelise seose tugevuse mõõtmiseks kasutatavat statistikut, mida olen testinud murdekorpuse ühendverbide peal sihiga välja selgitada parimad statistikud murdekorpuse ühendverbide tuvastamiseks ning murdekorpuse materjaliga töötamiseks laiemalt edaspidi. Statistika terminoloogias lähtun Rahvusvahelise Statistika Instituudi kodulehel olevast terminipangast (ISI).

2. Statistiliste meetodite tööpõhimõtetest

2.1. Mis andmeid on vaja sõnadevahelise seose tugevuse mõõtmise statistikute kasutamiseks?

Sõnadevahelise seose tugevuse väärtused on kvantitatiivselt mõõdetavad ja nad iseloomustavad sõnadevahelise seose tugevust. Neid leitakse erinevate statistiliste valemite abil. Väärtuse tõlgendamise põhimõte on lihtne: kõrged väärtused osutavad sõnadevahelisele tugevale seosele, väiksed ja negatiivsed väärtused sellele, et sõnad pigem välistavad koosesinemisel üksteist.

Selleks, et püsiühendeid sõnadevahelise seose tugevuse väärtuse järgi leida, peaks kõigepealt piiritlema, mida sõnade koosesinemise all mõeldakse. Kas soovitakse leida ainult n-ö tõelisi püsiühendeid või lihtsalt seada sõnadevahelise seose tugevuse mõõdud mingile skaalale, tõmbamata seejuures kindlat piiri püsiühendite ja mitte-püsiühendite vahele. (Evert 2009)

Teiseks peab kindlaks tegema, kuidas püsiühendeid grupeerida: kas ollakse huvitatud ainult väga tugeva seosega sõnaühenditest, mida peetakse täiesti iseseisvateks leksikaalseteks või grammatilisteks üksusteks, või soovitakse ainult kindlaks teha, milline on vaadeldava sõna levinum kontekst. N-ö tõeliste püsiühendite kindlakstegemisel määrab väärtuste piiri uurija ise, otsustades, kui suur peab olema sõnadevahelise seose tugevuse väärtus, et neid veel koosesinevaks pidada. (Evert 2009)

[2] Informatsiooniteooria on teadusharu, mis tegeleb läbi kanali saadetava info mõõtmise ja ülekandmise uurimisega (http://virtual01.lncc.br/~giraldi/qc/itheory.html) (01.09.2009).

Kõige lihtsam viis korpusest püsiühendeid leida on lihtsalt nad kokku lugeda – leida sõnade koosesinemise arv. Sellise lihtsa meetodiga võib saavutada juba üsna head tulemused ja eriti hästi töötab see läbipaistvate ning kindla konstruktsiooniga püsiühendite peal (*kõnet pidama*, *tantsu lööma*). Sageli võivad püsiühendi eri osad paikneda lauses üsna vabalt ning sellistel juhtudel jääb sageduste kokkulugemisest väheks, siin on vajalik seose tugevuse üle otsustamiseks keerulisemaid meetodeid. (Manning, Schütze 2003: 153–158)

Sageli jääb sõnadevahelise seose üle otsustamisel lihtsalt nende koosesinemiskordade kokkulugemisest väheks. Kui moodustada korpuse bigrammide ehk sõnapaaride sagedusloend, siis ilmselt oleks seal üsna kõrgel kohal ühend *ja ka*. See aga ei tähenda veel, et tegemist on püsiühendiga. Mõlemad sõnad esinevad korpuses sageli ning seega on loomulik, et ka nende koosesinemist tuleb tihti ette. Lisaks lihtsale koosesinemise sagedusele tuleb arvesse võtta ühendit moodustavate sõnade eraldiesinemise sagedusi – marginaal- ehk ääresagedusi.[3] Samuti peab arvestama korpuse suurust ehk valimimahtu, kust püsiühendeid leitakse. Seega saame statistiliste meetodite kasutamiseks sageduse arvutamise andmestiku, mis moodustub järgmistest näitajatest: O – sõnade koosesinemise sagedus korpuses ehk valimis, f1 ja f2 – vastavalt esimese ja teise sõna eraldiesinemise sagedus, N – valimimaht (valimimahust tuleb täpsemalt juttu edaspidi). (Evert 2009)

Lisaks on statistiliste meetodite kasutamiseks vajalik leida koosesinemise teoreetiline sagedus E (ingl *expected frequency*), mis näitab, kui suur on sõnade koosesinemise oodatav tõenäosus, eeldusel et sõnad esinevad tekstis üksteisest sõltumatult. Üks võimalus on arvutada teoreetiline sagedus järgmise valemi põhjal: E = f1*f2/N ehk sõnade eraldiesinemise sageduste korrutis tuleb jagada valimi suurusega. (Evert 2009)

Keerulisemate statistiliste meetodite jaoks on eelnevalt vaja koostada kahemõõtmeline sagedustabel (*contingency tabel*), mis arvutatakse sõnade eraldi- ja koosesinemissageduste ning valimimahu põhjal. Tabel 1 esitab kahemõõtmelise sagedustabeli ning tehted, kuidas leitakse väärtused tabeli eri lahtrite jaoks.

Tabel 1. Kahemõõtmeline sagedustabel (Evert 2009)

O_{11}	O	$f_1 - O$	O_{12}
O_{21}	$f_2 - O$	N – f1 – f2 + O	O_{22}

Oletame, et meil on korpus 48 705 osalausega (N = 48 705). Korpuses esinevad sõnad *vastu* ja *võtma* samas osalauses 78 korda (O = 78). *Vastu* ja *võtma* eraldiesinemise sagedused on vastavalt 458 ja 670 ehk f1 = 458 ja f2 = 670. Sellistel tingimustel saab koostada tabelis 2 esitatud kahemõõtmelise sagedustabeli.

Tabel 2. Kahemõõtmeline sagedustabel (*vastu võtma*)

O_{11}	78	458 – 78 = 380	O_{12}
O_{21}	670 – 78 = 592	48 705 – 458 – 670 + 78 = 47655	O_{22}

[3] Statistiline termin sõnade eraldiesinemissageduste kohta selles kontekstis on marginaal- ehk ääresagedused. Selles artiklis kasutan siiski terminit *sõnade eraldiesinemissagedus*, mis on intuitiivselt arusaadavam.

2.2. Kolm lähenemisviisi materjali liigendamiseks

Selles osas tutvustan kolme lähenemisviisi, millest võib lähtuda püsiühendite kandidaatpaaride moodustamisel. Näidetena olen siin kasutanud kirjakeelset materjali,[4] sest kirjakeeles on laused kirjavahemärgistatud ning laused ja osalaused selgesti eraldatavad, mistõttu on kirjeldatavaid põhimõtteid lihtsam mõista.

1. **Kindlas naabruses koosesinemine** (ingl *surface cooccurrence*) on kõige levinum lähenemine sõnadevahelise seose tugevuse mõõtmiseks. Selle järgi peetakse kahte sõna koosesinevaks, kui nad esinevad sageli koos teatavas kauguses ehk samas aknas (*collocational span*), mida arvestatakse kahe sõna vahele jäävate sõnede järgi. Akna suuruse määrab uurija, tavaliselt jääb see kolme ja viie sõna vahele. Akna suurust võib märkida kujul *L5* ja *R5*, mis tähendab, et uuritava sõna vasakule ning paremale poole arvestatakse viis sõna ehk aken on sümmeetriline. Aken võib olla ka asümmeetriline, näiteks *L2* ja *R4*, kui uuritava sõna vasakule poole jääb kaks ning paremale neli sõna. Samuti peab otsustama, kas arvestatakse ainult sõnu või ka muid märke (kirjavahemärke, numbreid). Tuleb määratleda, kuidas käsitletakse mitmesõnalisi ühendeid ning kas koosesinevad sõnad võivad ületada lausepiire. (Evert 2009)

Näide (1) illustreerib kindlas naabruses koosesinemist. Vaadeldavateks sõnadeks on *juttu* ja *ajama*. Uuritav sõnavorm on *juttu*. Akna suurus on *L3* ja *R3* ehk konteksti arvestatakse mõlemalt poolt kolm sõna. Osalause piire pole arvesse võetud, kirjavahemärke akna suuruse määramisel ei arvestata.

(1) Kiusu pärast, oma täieliku ükskõiksuse pärast tegin sellest siiski **juttu**, puudutasin keelatud asju. Viieaastaselt oli poiss leiutanud viisi, kuidas linnas **juttu** *ajama* jäänud õdesid kähku liikuma saada. Ta võis vabalt ka kogu selle aja vaikselt uksepeal nende **juttu** pealt kuulata ja seejärel tuppa astudes teatrit tegema hakata. Aga seal, emaga köögis **juttu** *ajades*, olevat tulnud talle just Alfredi tublidus meelde. Need mehed aga, kes niisugust hinnapoliitikat ***ajavad***, oma nägu kail ei näita, nemad istuvad kaldal uhketes büroomajades ning nende juurde iga soovijat ei lastagi.

Juttu ajama esineb samas aknas kaks korda ehk O = 2. Sõnade *juttu* ja *ajama* eraldiesinemise sagedused on vastavalt neli ja kolm ehk f1 = 4 ja f2 = 3. Valimimaht ehk N antud näite puhul N = 84 (kirjavahemärke pole arvestatud). Sõnade arvu järgi arvestatakse valimimahtu ainult kindlas naabruses koosesinemise lähenemisest lähtudes.

2. **Tekstilise koosesinemise** (*textual cooccurrence*) puhul peetakse sõnu koosesinevaks, kui nad esinevad koos samas tekstiüksuses, tavaliselt lauses, osalauses või lausungis. Kindlas naabruses koosesinemise lähenemisele heidetakse ette akna suuruse suvalisust, mis võib tulemust vales suunas mõjutada just vaba sõnajärjega keeltes, kus omavahel tugevas seoses olevad sõnad võivad üksteisest paikneda kaugel. Seega sobib sellistes keeltes paremini aknaks kogu lause või osalause ehk püsiühendite statistilisel tuvastamisel tuleb aluseks võtta tekstilise koosesinemise lähenemine. See lähenemine suudab tuvastada ka nõrgemas seoses olevaid püsiühendeid. (Evert 2009)

Tabel 3 illustreerib tekstilist koosesinemist. Siin võetakse arvesse osalause piire ning vaadeldakse sõnade *peale* ja *käima* koosesinemist samas osalauses.

[4] Eesti kirjakeele korpus: http://www.cl.ut.ee/korpused/kasutajaliides/ (30.08.2009).

Tabel 3. Tekstiline koosesinemine

Osalause	Info koosesinemise kohta
Ta oli ise riskialdis ja hasartne,	__ __
temal **käis** kõik ikka kogu panga **peale**.	peale käima
Selle vastu polnud mul kõige väiksematki huvi,	__ __
aga et ma olin Aimest kuulda saades kindlalt otsustanud Liisuga jälle sõber olla,	__ __
siis **käisin** talle tormiliselt **peale**:	peale käima
"Tule ikki!	__ __
Joost oleks varanduse ära andnud,	__ __
aga õpilane ei tahtnud vist,	__ __
või kui tahtiski,	__ __
siis ei julgenud **peale käia**,	peale käima
oli võltsilt viisakas, korralik.	__ __
Juhangi läks selle **peale** ägedaks.	peale
Eile õhtul **käisin** auhinda vastu võtmas.	käima

Ühend *peale käima* esineb koos samas osalauses kolm korda, seega O = 3. Sõnade eraldiesinemissagedused arvestatakse selle järgi, mitmes osalauses kumbki sõna esineb, antud näites f1 = 4 ja f2 = 4. Osalauseid on 13 ehk N = 13.

3. **Süntaktilise koosesinemise** (*syntactic cooccurrence*) lähenemine on kõige suuremate piirangutega, mille kohaselt sõnu peetakse koosesinevateks ainult juhul, kui nende vahel on kindel süntaktiline seos, näiteks verb koos subjekti või objektiga. Süntaktilise seose seisukohalt erinevaid püsiühendeid vaadeldakse eraldi. Süntaktilise koosesinemise lähenemine on sobilik olukordades, kus koosesinevad sõnad asuvad teineteisest kaugel. See lähenemine võetakse sageli aluseks püsiühendite tuvastamisel, kuna paljud leksikaliseerunud ühendid esinevad koos tavaliselt kindlas süntaktilises seoses (näiteks adverbi ja verbi ühend võib koos moodustada öeldise). (Evert 2009)

Süntaktilise koosesinemise lähenemist illustreerib tabel 4. Oletame, et soovime teada saada, kas *tantsu lööma* on püsiühend.

Tabel 4. Süntaktiline koosesinemine

Lause	Info koosesinemise kohta
Igatsedes vaatas[pred] Grethel noori[obj], kes jõululaule[obj] lausid[pred] ja lõbusalt **tantsu**[obj] **lõid**[pred].	vaatas noori laulsid jõululaule lõid tantsu
Koerad **lõid**[pred] **tantsu**[obj]**.**	lõid tantsu
Ta sulges[pred] silmad[obj] ja alustas[pred] **tantsu**[obj].	sulges silmad alustas tantsu
Ta **lõi**[pred] mind[obj] selgeks.	lõi mind
Kas te **lõite**[pred] mäataste vahel siis **tantsu**[obj]**?**	lõite tantsu
Ta avas[pred] **ukse**[obj].	avas ukse
Tantsisime[pred] kolm **tantsu**[obj].	tantsisime tantsu
Nad piinasid[pred] **tantsu ajal** jalgu[obj].	piinasid jalgu

Siin leitakse vajalikud andmed natuke erineval viisil. Oletame, et näites kasutatud väike korpus on süntaktiliselt märgendatud, s.t iga lauseliige on tähistatud vastava märgendiga. Antud juhul uurime ainult predikaadi ja objekti kombinatsioone, seega tuleb esmalt tuvastada kõik sellises seoses olevad sõnaühendid, mis koos moodustavad valimi. Tabelis 4 on 11 predikaadi ja objekti ühendit, seega N = 11. Eraldame uuritava ühendi *tantsu lööma*, näites on neid kolm, seega O = 3. Sõnade eraldiesinemissagedused on vastavalt f1 = 5 ning f2 = 4. Tegelikult esineb sõna *tantsu* tekstis 6 korda. Viimases lauses aga pole ta objekti rollis ning kuna antud juhul on huvi keskmes vaid objekti ja predikaadi ühendid, siis muus funktsioonis esinevaid vaatlusaluseid sõnavorme sageduste kokkulugemisel ei arvestata.

Kindlas naabruses koosesinemise lähenemine on osutunud kasulikuks leksikograafias ja korpuslingvistikas. Tekstilise koosesinemise lähenemist on kergem rakendada ning see on tõhusam kõikvõimalike juhuslike väljendite välistamisel. Süntaktilise koosesinemise lähenemine on tõhusam kindla konstruktsiooniga püsiühendite leidmisel, aga seab eelduseks korpuse süntaktilise märgendatuse.

3. Materjal

Käesoleva töö materjaliks on Eesti murrete korpuse morfoloogiliselt märgendatud tekstid (seisuga 01.06.2009) ning nendest leitud ühendverbid. Tekstid on eelnevalt lausestatud, kasutades Kaili Müürisepa loodud suulise kõne süntaksi analüsaatorit, mida on kohandatud murdetekstide lausestamise jaoks (selle kohta vt Lindström, Müürisep 2009). Lausestatud tekstide põhjal moodustasin kaheliikmelised kombinatsioonid kõigi osalause sõnadega, neist otsisin automaatselt välja adverbi ning verbi ühendid. Vaatluse all on ainult kaheliikmelised ühendid. Välja on jäetud verbi *olema* sisaldavad ühendid, samuti ei ole ma kaasanud modaalverbe (*saama*, *võima*, *tohtima*, *pidama*) sisaldavaid ühendeid, kuna need käituvad sageli abiverbina. Kui sõnakombinatsioonid moodustada osalause sees kõigi verbide ning adverbidega, moodustuks ühendverbi kandidaat ka abiverbi funktsioonis esineva verbivormiga, nt osalauses *töö sai ära tehtud* moodustuksid ühendid *ära tegema* kui *ära saama*, s.t toimuks teatav ülegenereerimine.[5]

Saadud verbi ja adverbi ühendite peal on lisaks rakendatud adverbide stoppsõnade loendit, s.t eemaldatud on ühendid, mille adverbiline komponent reeglina ühendverbi ei moodusta, nt *hullusti*, *jõudsalt*, *korrapäraselt* jt (täpsemalt vt Uiboaed 2008). Samuti olen analüüsist välja jätnud ühendid, mis esinevad samas osalauses vähem kui kolm korda.

Kõik katsed on tehtud kolmes osas, arvestades kolme erinevat murderühma. Põhjaeesti murderühma moodustavad idamurre, keskmurre, läänemurre ja saarte murre (kokku 279 126 sõna). Lõunaeesti murderühm hõlmab Mulgi, Tartu, Setu ja Võru murde (147 666 sõna). Madalaima esindatusega (75 144 sõna) rannikumurde murderühma moodustavad Alutaguse ning rannamurre.

Leidmata jäänud ühendverbide väljaselgitamiseks olen kasutanud testkorpust, mis sisaldab 10% kogu korpuse tekstidest. Rannikumurdes jääb testkorpuse põhjal leidmata 2% olemasolevatest ühendverbidest, lõuna- ja põhjaeesti rühmas on osakaalud mõlemad 16%. Korpuses olemasolevad ühendverbid jäävad leidmata peamiselt lausestaja vigade tõttu, sest ühendverbe leitakse ainult ühe osalause

[5] Uiboaed (2010) kirjeldatud katse, mis põhineb samal materjalil, erineb siinsest selle poolest, et seal pole ühtegi statistikut kasutatud: vastavate ühendverbide sagedused on lihtsalt kokku loetud (mitu korda nad ühes osalauses koos esinevad). Seega pole arvesse võetud ka sõnade eraldiesinemise sagedusi. Samuti olid seal osalausestamise alused teised (automaatsel osalausestamisel ei kasutatud süntaksianalüsaatorit).

piirides (piirid on seatud automaatselt). Kui lausestaja on osalause piiri seadnud valesti, nii et ühendverbi eri osad jäävad kahte erinevasse osalausesse, siis sellised ühendverbid jäävad tuvastamata.

4. Sõnadevahelise seose tugevuse mõõtmise statistikud

Selles peatükis näitan katsete abil, kuidas sõnadevahelise seose tugevuse mõõtmise statistikuid töötavad ning milliseid otsuseid nende põhjal teha võiks. Lisaks väga lihtsatele meetoditele (nt sageduste kokkulugemine) on olemas ka terve hulk informatsiooniteoorial ja statistilise olulisuse testimisel põhinevaid statistikuid. Mina olen tutvustamiseks välja valinud neli: kaks lihtsamat statistikut – 1) t-skoori (ingl *t-score*) ja 2) vastastikuse informatsiooni (MI) mõõdiku (*Mutual Information*) – ning kaks statistilist statistikut – 3) hii-ruut-statistiku (*chi-squered statistic*) ja 4) log-tõepära funktsiooni (*log-likelihood*).

Tavaliselt kasutatakse statistikuid kandidaatandmete seast tugevama seosega ühendite väljavalimiseks. Kõigepealt leitakse mingi hulk ühendeid (mõne osas 2.2 kirjeldatud meetodi abil) ning nende jaoks arvutatakse statistilise seose tugevuse väärtused. Seejärel järjestatakse loend ümber ning koosesinevateks loetakse ainult teatav hulk ühendeid sagedusloendi esimesest osast (väärtuse, millest alates ühendeid enam koosesinevateks ei loeta, peab uurija määrama ise). (Krenn, Evert 2001)

Arvutuste tegemisel lähtusin tekstilise koosesinemise põhimõttest – aknaks määrasin osalause, seega valimimaht on osalausete arv. Statistikute väärtused on arvutatud järgneva andmestiku abil: osalausete arv, ühendit moodustavate sõnade sagedused korpuses, sõnade koosesinemise arv samas osalauses ning nende koosesinemise oodatav sagedus. Kõik katsed on tehtud kolmes osas, arvestades kolme murderühma, seega iga murderühma sagedusandmed on erinevad. Iga statistiku kirjelduse juures olevad pingeread on järjestatud vastava statistiku väärtuste alusel, s.t eespool on ühendid mida statistik loeb tugevasti kokkukuuluvateks.

Statistikute töö kontrollimiseks vaatasin läbi sada kõrgema väärtusega ühendverbi igas murdrühmas ja iga statistiku kaupa eraldi. Saja kõrgema väärtusega ühendi hulgas on korpuses ühendverbidena mitteesinevaid ühendeid ühe ja kolme ühendi vahel ning enamjaolt on need ühendid sattunud loendisse märgendusvigade tõttu. Seega võib statistikute tööd pidada tõhusaks.

4.1. t-skoor

t-skoor arvutatakse järgmise valemi põhjal:

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

kus O on osalausete hulk, kus sõnad koos esinevad, ja E on teoreetiline sagedus.

Tabel 5 esitab kahanevalt 50 kõrgeima t-skoori väärtusega ühendit.

Tabel 5. 50 kõrgeima t-skoori väärtusega ühendit eri murderühmades
(k-sag – ühendi koosesinemissagedus samas osalauses, t-sk – t-skoori väärtus)

PÕHJA	k-sag	t-sk	LÕUNA	k-sag	t-sk	RANNIKU	k-sag	t-sk
peale_panema	97	8,59	ära_koolema	56	7,03	ära_võtma	30	4,42
välja_tulema	92	8,11	peale_panema	37	5,52	maha_laskma	19	4,19
sisse_panema	91	7,82	ära_surema	35	5,44	peale_hakkama	20	4,03
ära_võtma	103	7,50	sisse_panema	31	4,83	välja_vedama	20	4,03
ära_surema	63	7,39	ära_tapma	22	4,21	sisse_laskma	16	3,80
ära_kaduma	51	6,55	välja_tulema	22	3,75	tagasi_tulema	18	3,65
ära_viima	68	6,37	maha_jääma	16	3,70	sisse_panema	17	3,55
valmis_tegema	42	5,61	üles_ajama	15	3,65	ära_surema	16	3,53
üles_panema	50	5,55	tagasi_tulema	16	3,54	välja_tulema	24	3,30
kinni_siduma	31	5,47	üles_tulema	20	3,53	maha_lööma	12	3,29
ära_tooma	57	5,43	vastu_tulema	14	3,37	ringi_käima	13	3,29
alla_panema	39	5,42	kinni_hoidma	12	3,36	ära_lahutama	11	3,14
lahti_võtma	34	5,38	kinni_panema	18	3,36	ära_tulema	34	3,12
välja_võtma	43	5,34	ära_minema	65	3,34	läbi_käima	14	3,09
maha_laskma	29	4,99	ära_kuivatama	12	3,28	ära_viima	15	3,04
läbi_käima	37	4,95	kinni_võtma	15	3,27	kinni_panema	13	2,94
ära_tapma	29	4,88	ära_müüma	13	3,19	vastu_võtma	10	2,93
ette_panema	31	4,84	ühes_võtma	12	3,17	välja_minema	22	2,83
ära_kuivatama	26	4,63	kinni_köitma	10	3,12	välja_võtma	13	2,81
kinni_panema	38	4,58	ette_panema	11	3,06	vastu_tulema	11	2,72
tagasi_tulema	25	4,39	üles_panema	16	3,01	peale_panema	11	2,67
vastu_võtma	19	4,11	ära_uppuma	10	2,96	lahti_tegema	11	2,65
sisse_minema	43	4,10	vastu_võtma	10	2,91	üles_võtma	10	2,65
kokku_vedama	18	4,08	välja_võtma	13	2,90	alla_vedama	8	2,59
ära_sööma	32	4,00	sisse_tulema	16	2,90	üles_tõstma	7	2,58
sisse_laskma	22	3,97	ära_pesema	12	2,90	ette_panema	8	2,51
pealt_võtma	17	3,84	manu_panema	10	2,80	valmis_tegema	8	2,48
üles_tõusma	15	3,80	ära_võtma	33	2,77	edasi_minema	10	2,48
maha_kukkuma	14	3,67	välja_ajama	9	2,70	üles_panema	11	2,47
välja_minema	41	3,67	ära_katkuma	10	2,67	ära_kaduma	7	2,46
läbi_pistma	14	3,67	ära_viima	17	2,59	maha_kukkuma	6	2,43
ära_kuivama	17	3,63	maha_tapma	7	2,53	lahti_päästma	6	2,43
peale_hakkama	25	3,62	ära_peksma	12	2,51	kinni_siduma	6	2,41
sisse_ajama	19	3,59	ära_keetma	10	2,49	üles_tõusma	6	2,41
ära_müüma	17	3,53	ära_eksima	7	2,48	mööda_minema	7	2,38
segamini_ minema	16	3,53	ära_lõikama	11	2,46	kokku_ajama	6	2,35
ära_murdma	15	3,50	sisse_lööma	7	2,45	läbi_ajama	6	2,32

PÕHJA	k-sag	t-sk	LÕUNA	k-sag	t-sk	RANNIKU	k-sag	t-sk
maha_panema	29	3,46	ära_kaduma	8	2,45	külge_panema	6	2,31
ära_kaotama	13	3,44	katki_lööma	6	2,41	välja_paistma	6	2,30
ülal_istuma	12	3,43	välja_minema	15	2,41	ära_vajuma	6	2,28
maha_võtma	20	3,43	üle_minema	8	2,41	tagasi_minema	10	2,27
ära_lõhkuma	14	3,43	ühes_käima	7	2,37	välja_viima	8	2,26
välja_ajama	18	3,39	taga_ajama	6	2,37	püsti_tõstma	5	2,23
välja_laskma	18	3,38	ära_hõõruma	7	2,29	kokku_vedama	6	2,22
ära_tulema	77	3,34	ära_kuivama	7	2,29	sisse_lööma	6	2,19
üles_ajama	15	3,31	sisse_laskma	7	2,28	kokku_korjama	5	2,19
ära_lõikama	19	3,27	peale_valama	5	2,20	maha_jätma	5	2,18
üles_võtma	19	3,26	mööda_ minema	6	2,20	üles_leidma	5	2,17
kokku_panema	22	3,25	maha_sadama	5	2,19	ära_rookima	5	2,15
ära_niitma	17	3,25	valla_laskma	5	2,19	ära_korjama	6	2,13

t-skoor järjestab ühendverbid kõigis murderühmades suhteliselt hästi ehk sagedamad ühendverbid saavad kõrgema t-skoori väärtuse. Kuigi analüüsist on välja jäetud *saama*-ühendid, võis t-skoori väärtuste põhjal näha, et vastavad ühendid saavad suhteliselt madala väärtuse, mis vastab tegelikkusele, sest tegemist ei ole ühendverbide vaid adverbi ja abiverbi funktsioonis esineva verbivormi ühenditega. Eriti hästi oli seda näha põhjaeesti murderühmas, kus kasutatakse palju *saama*-passiivi. Sama kehtib paljude *pidama*-ühendite kohta nii põhja- kui lõunaeesti murderühmas.

Kuigi Stefan Evert on oma doktoritöös t-skoori kritiseerinud, pidades seda täiesti sobimatuks statistikuks sõnadevahelise seose mõõtmiseks (Evert 2004: 82–83), on ta teinud katseid, kus t-skoor on andnud häid tulemusi (Krenn, Evert 2001). Sama näitab ka murdematerjal.

t-skoori väärtuste põhjal tuleb selgelt välja statistikute kasutamise suurim eelis lihtsalt sageduste kokkulugemise ees – statistik võimaldab arvesse võtta sõnade eraldiesinemise sagedusi. Sellest tulenevalt saavad madalama väärtuse sagedastest sõnadest moodustunud ühendid, mis ei moodusta reeglina ühendverbi (*välja pidama*, *alla pidama*, *ära hakkama*). t-skoori tulemuste parandamiseks oleks vaja meetodit, mis välistaks sagedatest sõnadest moodustunud ühendverbide mitte väljajäämist vaatluse alt. Nii nagu on juhtunud lõunaeesti murderühmas, kus madala väärtuse on saanud mõned sagedatest sõnadest moodustunud ühendverbid (*ära tegema*, *sisse saama*). Üks võimalus selleks on kõrge koosesinemissagedusega ja madalate t-skoori väärtustega ühendid käsitsi läbi vaadata, mis küll praeguse murdekorpuse mahu juures oleks mõeldav, kuid kindlasti mitte efektiivseim meetod.

Murdematerjali t-skoori väärtuse piiri, millest väiksema väärtusega ühendeid enam kokkukuuluvateks ei loeta, pole kerge seada. Osaliselt on see kindlasti tingitud murdekorpuse materjali vähesusest, sest üldjuhul töötavad statistikud paremini väga suuremahuliste andmekogude peal. Nende katsete põhjal võiks ranniku- ja põhjaeesti murderühmas seada t-skoori piirväärtuseks 1 ning lõunaeesti murderühmas 0,5, s.t kui t-skoor on väiksem ühest või 0,5-st, siis ühendit enam ühendverbiks ei loeta.

4.2. Vastatikuse informatsiooni väärtus (MI)

Vastastikuse informatsiooni väärtus MI leitakse valemiga:

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

kus O on osalausete hulk, kus sõnad koos esinevad, ja E on teoreetiline sagedus.

Tabel 6 esitab kahanevalt 50 kõrgeima MI väärtusega ühendit.

MI eelistab harvaesinevaid ühendeid, mille eri osad on samuti korpuses madala esinemissagedusega (Evert 2009). Seda kinnitavad ka minu katsed. Everti järgi osutab negatiivne MI väärtus antikollokatsioonidele ehk sõnadele, mis pigem välistavad teineteise läheduses esinemise.

Põhjaeesti murderühmas said kõrgema negatiivse väärtuse pigem sageli koosesinevad sõnaühendid (*ära tegema*, *ära ütlema*, *ära käima*, *ära panema*) ning kontrolli käigus selgub, et tõepoolest pole tegemist sagedaste ühendverbidega. Kõrgem koosesinemissagedus tuleneb asjaolust, et süntaksianalüsaator pole hakkama saanud korrektse osalausestamisega, mistõttu on üheks osalauseks loetud mitu ning sellest tulenevalt on ühendverbi kandidaate ülegenereeritud. Statistik ise ei saa arvesse võtta, kas tegelikkuses on tegemist ühe või mitme osalausega. Lõunaeesti murderühmas saavad samalaadsed ühendid samuti kõrge negatiivse väärtuse, kuid antud murderühmas on tegemist siiski ühendverbidega, mis tähendab, et statistiku väärtus ei peegelda tegelikkust. Rannikumurde murderühmas sarnaselt põhjaeesti murderühmaga on küll paljud mitte-ühendverbid saanud madala väärtuse, kuid kõrge koha pingereas on saanud näiteks *välja varastama*, *pealt jääma*, *maha lõikama*, *välja lõppema*, mis ei esine tekstis ühendverbina. Seega on MI käitumine eri murderühmades mõnevõrra erinev.

Tabel 6. 50 kõrgeima MI väärtusega ühendit eri murderühmades
(k-sag – ühendi koosesinemissagedus samas osalauses, MI – MI väärtus)

PÕHJA	k-sag	MI	LÕUNA	k-sag	MI	RANNIKU	k-sag	MI
segamini_ vaduma	3	9,16	mööda_sõitma	4	7,85	püsti_tõstma	5	8,13
ümbert_siduma	3	8,57	peale_siputama	3	7,51	maha_kukkuma	6	7,30
katki_närima	3	7,65	ümber_mähkima	3	7,22	lahti_päästma	6	6,69
alla_neelama	3	7,41	maha_murduma	3	6,96	üles_lendama	3	6,18
kinni_sõlmima	3	7,24	kinni_haarama	3	6,95	tagasi_langema	3	6,12
ilma_jätma	3	7,07	õkva_tõmbama	3	6,80	kinni_siduma	6	5,99
kinni_katma	7	6,87	kinni_kängima	4	6,78	üles_tõusma	6	5,86
ülal_istuma	12	6,87	üles_atma	4	6,49	kokku_korjama	5	5,48
seest_kiskuma	3	6,81	üles_tõusma	4	6,49	katki_lõikama	3	5,44
alla_vajuma	3	6,74	perra_jätma	3	6,48	ära_triivima	3	5,35
ringi_pöörama	3	6,63	kinni_katma	4	6,37	ära_täitma	3	5,35
hiljaks_jääma	4	6,55	kinni_köitma	10	6,23	ringi_keerama	4	5,34
üles_pooma	4	6,51	maha_kargama	4	6,06	üles_tõstma	7	5,29

PÕHJA	k-sag	MI	LÕUNA	k-sag	MI	RANNIKU	k-sag	MI
lahti_kaevama	3	6,24	katki_lööma	6	6,01	maha_jätma	5	5,26
üleval_istuma	3	6,19	peale_valama	5	5,92	üles_leidma	5	5,18
kinni_kargama	5	6,10	maha_röökima	4	5,79	edasi_lükkama	3	5,15
üles_paisutama	4	5,93	maha_sadama	5	5,61	kinni_hoidma	4	5,08
peksa_andma	6	5,89	valla_laskma	5	5,56	maha_müüma	4	5,00
edekohe_andma	3	5,89	kinni_mähkima	4	5,46	ära_selgima	4	4,76
ümber_pöörama	5	5,80	välja_kiskuma	5	5,39	ära_rookima	5	4,67
kinni_siduma	31	5,76	ilma_jääma	4	5,26	maha_laskma	19	4,65
üles_tõusma	15	5,76	kinni_hoidma	12	5,07	kokku_ajama	6	4,59
maha_vajuma	3	5,73	alla_lööma	5	5,03	kaasa_tooma	3	4,57
maha_kukkuma	14	5,71	tähele_panema	3	4,96	taga_vahtima	3	4,43
katki_lõikama	9	5,69	maha_viskama	5	4,95	taga_ajama	4	4,39
katki_raiuma	3	5,68	taga_ajama	6	4,93	maha_lööma	12	4,35
edasi_lükkama	3	5,68	peale_keema	4	4,92	ära_kaotama	4	4,35
läbi_pistma	14	5,63	maha_tapma	7	4,51	järele_jääma	3	4,33
välja_rapsama	3	5,63	pakku_minema	3	4,50	sisse_laskma	16	4,30
lõhki_lööma	4	5,55	ära_kustuma	4	4,49	ümber_lööma	4	4,27
lõhki_lõikama	3	5,48	ära_puherduma	4	4,49	läbi_ajama	6	4,24
kokku_laduma	4	5,38	ära_kuivatama	12	4,27	ära_lahutama	11	4,22
ilma_jääma	6	5,33	ära_hääduma	5	4,22	edasi_ajama	4	4,22
tagasi_lükkama	4	5,30	ära_jahtuma	5	4,22	külge_panema	6	4,16
sisse_soolama	3	5,26	ära_külmuma	5	4,22	välja_varastama	3	4,14
kinni_tarima	3	5,24	ära_rehkma	5	4,22	pealt_jääma	3	4,11
järgi_andma	7	5,19	ära_survima	5	4,22	välja_paistma	6	4,04
põiki_laskma	3	5,18	ära_vassima	5	4,22	tagasi_keerama	3	3,99
välja_noppima	8	5,04	vahele_panema	3	4,22	ära_uppuma	4	3,89
järele_jääma	11	5,01	üles_viskama	3	4,13	ära_vajuma	6	3,85
valmis_ õmblema	3	4,98	üles_ajama	15	4,11	ära_kaduma	7	3,83
välja_kannatama	3	4,95	ära_vahetama	3	4,07	vastu_võtma	10	3,77
maha_jätma	11	4,90	mant_tulema	4	4,05	eemale_mi-nema	4	3,76
püsti_seisma	3	4,87	ära_koolema	56	4,05	maha_lõikama	3	3,72
peale_laduma	5	4,82	läbi_ajama	5	4,05	välja_lõppema	3	3,70
kokku_lükkama	5	4,81	täis_sööma	4	4,03	peitu_minema	3	3,60
kokku_siduma	11	4,80	üles_käänama	3	4,02	sisse_ajama	5	3,59
ümber_keerama	3	4,78	ära_uppuma	10	4,00	maha_surema	5	3,58
vargil_käima	4	4,75	ära_leotama	5	4,00	alla_vedama	8	3,55
sisse_puurima	3	4,74	ära_väsima	5	4,00	ringi_käima	13	3,29

Selline tulemus on tugevasti mõjutatud asjaolust, et ühendit moodustavad sõnad on korpuses väga sagedad ning MI töö põhimõttele vastavalt peavadki nad madala väärtuse saama. MI väärtust mõjutab tugevasti ainult sõnade eraldiesinemise sagedus, nende koosesinemise sagedus samas osalauses ei paista tulemusele olulisest mõju avaldavat, nagu võib näha rannikumurde murderühmas. Rannikumurde murderühmas oli MI töö eriti tõhus, mis puudutas modaalverbi ühendeid – MI väärtuste viimase osa moodustasid suures osas just ainult adverbi ja modaalverbi ühendid, mis aitab välistada selliste ühendite arvestamist ühendverbide hulka.

Kuna MI on loodud madala esinemissagedusega ühendite leidmiseks, siis võimendab statistik selliste ühendite väärtusi üle, mistõttu satuvad nad pingereas kõrgetele kohtadele. Selle võimenduse tasakaalustamiseks on pakutud erinevaid lahendusi. Üks võimalus on kogu valem läbi korrutada sõnade koosesinemissageduse väärtusega (Evert 2009). Seda tehes muutub MI väärtuste loend märgatavalt – kõrgema MI väärtuse saavad sagedamini esinevad ühendverbid.

Teise võimalusena on välja pakutud ühendite koosesinemise sagedusele piiri seadmist, kõige efektiivsemaks on peetud kümmet, ehk sõnad, mis esinevad samas osalauses vähem kui kümme korda, jäetakse vaatluse alt välja (Evert 2009). Rakendades sama tehnikat Eesti murdematerjali peal, moodustub MI väärtuste pingerida n-ö väga tugevatest ühendverbidest, mis on kindlasti lähenemisviisi eelis. Antud materjali peal ei oleks siiski see lahendus, kuna suurem osa ühendeid esinebki korpuses samas osalauses vähem kui kümme korda, näiteks rannikumurde murderühmas jääks selle tulemusena alles vaid vähem kui 50 ühendit.

MI väärtuse üle, millest alates ühendit enam ühendverbiks ei peeta, pole kerge otsustada, Eesti murdekorpuse puhul oleks see ilmselt võimatu. Sellise väärtuse seadmine tuleks kõne alla, kui jätta välja näiteks vähem kui kümme korda esinevad ühendid ning järelejäänud ühenditele seada teatud piirväärtus. Nagu eelnevalt mainitud, poleks see siiski murdekorpust silmas pidades selle vähesuse tõttu ainumõeldav lahendus.

Eesti murdematerjali peal võiks MI tulemuse parandamiseks kombineerida erinevaid väljapakutud lahendusi. Näiteks võib esialgse kandidaatandmestiku põhjal leida madala esinemissagedusega ühendverbid, seades piiri, millest vaatluse alla jäävad ainult eespool olevad ühendid. Seejärel võiks järjestada esialgse loendi uuesti, jättes välja vähem kui kümme korda samas osalauses esinevad ühendid. Saadud loendile võiks rakendada samuti piirväärtuse, mis välistaks kandidaatide seast sagedasti samas osalauses koosesinevad mitte-ühendverbid.

Krenn ja Evert (2001) väidavad, et MI ei sobi üldse näiteks väljendverbide ja muude kujundlike püsiühendite leidmiseks. Murdekorpuse ühendverbide leidmisel on MI mõningate kitsenduse seadmisel suhteliselt efektiivne.

4.3. Hii-ruut-statistik

Hii-ruudu arvutamiseks kasutatakse kahemõõtmelist sagedustabelit (vt eespool) ja arvutatakse valemi

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

abil, kus O on osalausete hulk, kus sõnad koos esinevad, ja E on teoreetiline sagedus.

Kuna hii-ruut on n-ö kahepoolne mõõdik (s.t annab suure positiivse väärtuse nii tugevale seosele – sõnadele, mis kindlasti moodustavad püsiühendi kui ka tugevale "tõukumisele" – sõnadele, mis "väldivad" üksteise naabrust), jäädes kogu aeg positiivseks (ruudu väärtus on alati positiivne), peaks tulemuse läbi korrutama (–1)-ga, juhul kui O < E (teoreetiline koosesinemissagedus on suurem tegelikust koosesinemissagedusest, s.t sõnad esinevad üksteise naabruses harvem, kui võib oodata nende eraldiesinemise sageduste põhjal), saades hii-ruudust ühepoolse mõõdiku, mis eristab nii positiivseid kui negatiivseid seoseid (Evert 2009).

Tabel 7 esitab kahanevalt 50 kõrgeima hii-ruudu väärtusega ühendit (väärtused on läbi korrutatud (–1)-ga, juhul kui O < E).

Hii-ruut-statistik eelistab sarnaselt MI-ga madala esinemissagedusega sõnu ja ühendeid, samas antakse kõrged väärtused n-ö väga tugevatele ühendverbidele. Hii-ruudu väärtuste järjestus sarnaneb MI väärtuste loendiga. Madala väärtuse põhjaeesti murderühmas saavad sageli koosesinevad sõnad, mis üldjuhul ei ole ühendverbid (*välja pidama*, *välja saama*, *välja hakkama*).

Tabel 7. 50 kõrgeima hii-ruut statistiku väärtusega ühendit eri murderühmades
(k-sag – ühendi koosesinemissagedus samas osalauses, hii-ruut – hii-ruut statistiku väärtus)

PÕHJA	k-sag	hii-ruut	LÕUNA	k-sag	hii-ruut	RANNIKU	k-sag	hii-ruut
segamini_ vaduma	3	1707,98	mööda_ sõitma	4	913,86	püsti_tõstma	5	1391,25
kinni_siduma	31	1638,07	ära_koolema	56	858,84	maha_ kukkuma	6	939,95
ülal_istuma	12	1384,91	kinni_köitma	10	736,05	lahti_ päästma	6	610,66
ümbert_ siduma	3	1138,12	peale_ siputama	3	543,91	maha_ laskma	19	447,34
ära_surema	63	823,39	ümber_ mähkima	3	440,82	kinni_ siduma	6	371,02
kinni_katma	7	812,33	kinni_ kängima	4	435,17	üles_tõusma	6	339,68
üles_tõusma	15	787,28	ära_surema	35	387,63	sisse_laskma	16	288,22
maha_ kukkuma	14	708,43	kinni_hoidma	12	382,88	üles_tõstma	7	262,42
läbi_pistma	14	671,67	katki_lööma	6	377,78	maha_lööma	12	224,82
peale_panema	97	609,76	maha_ murduma	3	371,09	kokku_ korjama	5	214,02
katki_närima	3	595,87	kinni_ haarama	3	367,89	üles_ lendama	3	213,65
ära_kaduma	51	538,64	üles_atma	4	353,46	tagasi_ langema	3	203,66
alla_neelama	3	507,85	üles_tõusma	4	353,46	ära_ lahutama	11	188,83
katki_lõikama	9	450,56	peale_ panema	37	344,77	maha_jätma	5	183,11
kinni_sõlmima	3	449,24	õkva_ tõmbama	3	329,49	üles_leidma	5	173,28
välja_tulema	92	448,49	kinni_katma	4	324,42	peale_ hakkama	20	170,21
kokku_ vedama	18	439,72	peale_valama	5	295,72	välja_ vedama	20	167,42
ilma_jätma	3	399,20	perra_jätma	3	262,41	ringi_ keerama	4	154,45

PÕHJA	k-sag	hii-ruut	LÕUNA	k-sag	hii-ruut	RANNIKU	k-sag	hii-ruut
lahti_võtma	34	385,92	maha_ kargama	4	260,25	kokku_ ajama	6	133,90
hiljaks_jääma	4	372,01	maha_ sadama	5	236,67	ringi_käima	13	130,54
üles_pooma	4	359,69	üles_ajama	15	233,87	kinni_ hoidma	4	128,28
sisse_panema	91	358,14	valla_laskma	5	228,31	katki_ lõikama	3	124,43
peksa_andma	6	347,47	ära_ kuivatama	12	217,27	maha_ müüma	4	120,98
maha_laskma	29	346,67	maha_ röökima	4	215,58	ära_rookima	5	120,47
järele_jääma	11	337,66	välja_kiskuma	5	201,29	vastu_võtma	10	119,93
kinni_kargama	5	334,98	maha_jääma	16	183,30	ära_triivima	3	119,18
seest_kiskuma	3	331,07	sisse_panema	31	183,12	ära_täitma	3	119,18
alla_vajuma	3	315,19	ära_tapma	22	179,70	ära_surema	16	108,50
maha_jätma	11	310,54	taga_ajama	6	172,31	ära_võtma	30	105,34
vastu_võtma	19	301,97	kinni_ mähkima	4	169,39	ära_selgima	4	103,31
ringi_pöörama	3	291,95	alla_lööma	5	154,04	läbi_ajama	6	102,67
kokku_siduma	11	287,27	maha_tapma	7	147,61	edasi_ lükkama	3	100,78
ümber_ pöörama	5	269,52	ära_uppuma	10	147,59	tagasi_ tulema	18	99,62
ära_tapma	29	263,55	ilma_jääma	4	146,75	külge_ panema	6	98,58
ära_kaotama	13	261,44	maha_ viskama	5	145,46	sisse_ panema	17	93,25
välja_noppima	8	251,09	ette_panema	11	122,62	ära_kaduma	7	88,55
valmis_ tegema	42	248,17	vastu_tulema	14	120,12	välja_ paistma	6	88,17
järgi_andma	7	243,65	ühes_võtma	12	119,05	alla_vedama	8	79,48
ära_kuivatama	26	241,40	peale_keema	4	114,42	taga_ajama	4	76,77
üles_ paisutama	4	237,17	tagasi_ tulema	16	111,74	ära_vajuma	6	76,62
alla_panema	39	232,90	vastu_võtma	10	108,97	ära_kaotama	4	75,53
ilma_jääma	6	232,41	ära_eksima	7	100,95	ümber_ lööma	4	69,78
ära_võtma	103	229,70	ära_müüma	13	91,28	edasi_ajama	4	67,20
lahti_kaevama	3	221,43	tähele_ panema	3	90,50	kaasa_ tooma	3	66,14
pealt_võtma	17	219,66	ära_hääduma	5	87,70	mööda_ minema	7	60,70
üleval_istuma	3	213,94	ära_jahtuma	5	87,70	taga_ vahtima	3	59,02
tähele_ panema	11	212,68	ära_külmuma	5	87,70	ette_ panema	8	58,26
ära_sulatama	10	211,90	ära_rehkma	5	87,70	läbi_käima	14	57,26
ette_panema	31	188,77	ära_survima	5	87,70	järele_jääma	3	55,12
ära_viima	68	188,17	ära_vassima	5	87,70	ära_uppuma	4	52,84

Sarnaselt MI-ga ei piisaks ka hii-ruudu puhul kollokatiivsuse ja mitte-kollokatiivsuse piirväärtuse seadmisest, kuna nii oleks vaatluse alla jäävate ühendite loend suhteliselt kaootiline. Nii nagu MI puhulgi paraneks tulemus, kui välja jätta ühendid, mis esinevad koos samas osalauses vähem kui kümme korda. Sellisel juhul jällegi jääksid välja harvaesinevad ühendverbid ning kuna murdekorpuse materjal on suhteliselt väike, on paljude ühendverbide esinemissagedus väiksem kui kümme. Hii-ruut oleks murdematerjali peal tulemuslikum, kui välja jätta madala esinemissagedusega sõnadest koosnevad ühendid, sest erinevalt MI-st ei "lükka" ta pingerea lõppu kõrge esinemissagedusega ühendeid. Madala sagedusega ühendite tuvastamisel töötab paremini MI.

4.4. Log-tõepära funktsioon

Log-tõepära funktsiooni väärtuse leidmiseks kasutatakse samuti kahemõõtmelist sagedustabelit (vt eespool) ning arvutatakse valemi

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

abil, kus O on osalausete hulk, kus sõnad koos esinevad, ja E on teoreetiline sagedus.

Nii nagu hii-ruut on ka log-tõepära funktsioon kahepoolne mõõdik ning selle ühepoolseks muutmiseks tuleb statistiku väärtus läbi korrutada (–1)-ga, juhul kui O < E.

Tabel 8 esitab kahanevalt 50 kõrgeima log-tõepära funktsiooni väärtusega ühendit (väärtused on läbi korrutatud (–1)-ga, juhul kui O < E).

Kirjanduse põhjal võis eeldada, et parima tulemuse annab log-tõepära funktsiooni kasutamine ning Eesti murdematerjalil tehtud katsed kinnitasid seda. Log-tõepära funktsioon on loodud just arvutilingvistiliste rakenduste jaoks ja on leidnud palju kasutust selles valdkonnas. Seda peetakse sobivaks paljude erinevate ülesannete lahendamiseks (Evert 2009, Krenn, Evert 2001). Kõigis murderühmades peegeldab log-tõepära väärtuste loend kõige enam tegelikkust ning Everti järgi korreleeruvad log-tõepära funktsiooni väärtused kõige rohkem inimeste poolt antud hinnangutega. Samuti võis kõigis kolmes murderühmas näha, et log-tõepära annab pigem väiksemad väärtused *saama*-ühenditele, mis reeglina on abiverbi ja adverbi ühendid (eriti just põhjaeesti murderühmas). See osutab taas statistikute kasutamise eelisele lihtsalt ühendite sageduste kokkulugemise ees. Samuti ei võimenda log-tõepära funktsioon üle madala sagedusega andmestikku.

Log-tõepära funktsiooni väärtust, millest alates ühendeid enam koosesinevateks ei loeta, pole murdematerjali peal kerge määrata. Nagu eelnevalt mainitud, on üheks põhjuseks korpuse väiksus. Tehtud katsete põhjal võiks väärtuse piir, millest alates sõnu enam koosesinevateks ei loeta, olla põhjaeesti murderühmas umbes 13 miljardi kandis, lõunaeesti murderühmas 4,5 miljardi kandis ning kirderanniku murderühmas umbes 2 miljardit.

Tabel 8. 50 kõrgeima log-tõepära funktsiooni väärtusega ühendit eri murderühmades (k-sag – ühendi koosesinemissagedus samas osalauses, log – log-tõepära väärtus)

PÕHJA	k-sag	log	LÕUNA	k-sag	log	RANNIKU	k-sag	log
peale_ panema	97	2,15E+11	ära_koolema	56	7,21E+10	maha_laskma	19	1,25E+10
ära_surema	63	2E+11	ära_surema	35	3,6E+10	ära_võtma	30	9,68E+09
välja_tulema	92	1,77E+11	peale_ panema	37	3,29E+10	sisse_laskma	16	9,54E+09
sisse_panema	91	1,61E+11	sisse_panema	31	2,24E+10	peale_ hakkama	20	9,42E+09
ära_kaduma	51	1,43E+11	ära_tapma	22	1,91E+10	välja_vedama	20	9,12E+09
ära_võtma	103	1,38E+11	ära_ kuivatama	12	1,81E+10	maha_ kukkuma	6	7,75E+09
kinni_siduma	31	1,31E+11	kinni_köitma	10	1,61E+10	tagasi_tulema	18	7,32E+09
ära_viima	68	1,01E+11	üles_ajama	15	1,48E+10	maha_lööma	12	7,16E+09
valmis_ tegema	42	9,03E+10	kinni_hoidma	12	1,48E+10	ära_lahutama	11	6,98E+09
lahti_võtma	34	8,95E+10	maha_jääma	16	1,43E+10	ringi_käima	13	6,88E+09
alla_panema	39	8,34E+10	tagasi_tulema	16	1,25E+10	ära_surema	16	6,8E+09
üles_panema	50	7,75E+10	ära_uppuma	10	1,24E+10	sisse_panema	17	6,65E+09
maha_laskma	29	7,53E+10	välja_tulema	22	1,23E+10	püsti_tõstma	5	5,99E+09
ära_tapma	29	7,51E+10	vastu_tulema	14	1,2E+10	lahti_päästma	6	5,86E+09
ära_tooma	57	7,13E+10	ära_minema	65	1,16E+10	vastu_võtma	10	5,21E+09
välja_võtma	43	7,1E+10	üles_tulema	20	1,07E+10	välja_tulema	24	5,19E+09
ära_ kuivatama	26	6,8E+10	ette_panema	11	1,05E+10	üles_tõstma	7	5,15E+09
ette_panema	31	6,68E+10	ära_müüma	13	1,05E+10	kinni_siduma	6	5,13E+09
üles_tõusma	15	6,29E+10	ühes_võtma	12	1,04E+10	üles_tõusma	6	5,06E+09
läbi_käima	37	6,17E+10	kinni_panema	18	9,7E+09	ära_tulema	34	4,96E+09
ära_kaotama	13	5,84E+10	kinni_võtma	15	9,56E+09	läbi_käima	14	4,87E+09
kokku_ vedama	18	5,82E+10	vastu_võtma	10	8,96E+09	ära_viima	15	4,45E+09
ülal_istuma	12	5,81E+10	katki_lööma	6	8,8E+09	kinni_panema	13	4,27E+09
maha_ kukkuma	14	5,77E+10	ära_eksima	7	8,5E+09	välja_minema	22	3,9E+09
vastu_võtma	19	5,71E+10	ära_pesema	12	7,78E+09	ära_kaduma	7	3,86E+09
läbi_pistma	14	5,66E+10	manu_ panema	10	7,73E+09	külge_ panema	6	3,81E+09
tagasi_tulema	25	5,46E+10	mööda_ sõitma	4	7,71E+09	alla_vedama	8	3,81E+09
tähele_ panema	11	5,33E+10	üles_panema	16	7,62E+09	vastu_tulema	11	3,77E+09
kinni_panema	38	5,12E+10	kinni_ kängima	4	7,51E+09	ära_rookima	5	3,77E+09
ära_sulatama	10	4,87E+10	maha_tapma	7	7,5E+09	välja_võtma	13	3,75E+09
pealt_võtma	17	4,69E+10	peale_valama	5	7,34E+09	kokku_ korjama	5	3,71E+09
sisse_minema	43	4,08E+10	ära_hääduma	5	7,25E+09	kokku_ajama	6	3,7E+09

PÕHJA	k-sag	log	LÕUNA	k-sag	log	RANNIKU	k-sag	log
kinni_katma	7	4,07E+10	ära_jahtuma	5	7,25E+09	mööda_ minema	7	3,62E+09
sisse_laskma	22	4E+10	ära_külmuma	5	7,25E+09	maha_jätma	5	3,59E+09
järele_jääma	11	3,96E+10	ära_rehkma	5	7,25E+09	üles_leidma	5	3,57E+09
ära_lõhkuma	14	3,86E+10	ära_survima	5	7,25E+09	ette_panema	8	3,52E+09
ära_murdma	15	3,85E+10	ära_vassima	5	7,25E+09	üles_võtma	10	3,49E+09
ära_sööma	32	3,85E+10	ära_võtma	33	7,23E+09	peale_ panema	11	3,48E+09
ära_kuivama	17	3,79E+10	välja_võtma	13	7,17E+09	lahti_tegema	11	3,46E+09
maha_jätma	11	3,73E+10	taga_ajama	6	7,13E+09	valmis_ tegema	8	3,4E+09
segamini_ minema	16	3,64E+10	sisse_tulema	16	7E+09	läbi_ajama	6	3,4E+09
kokku_ siduma	11	3,59E+10	välja_ajama	9	7E+09	ära_vajuma	6	3,32E+09
katki_lõikama	9	3,52E+10	valla_laskma	5	6,95E+09	välja_paistma	6	3,3E+09
ära_tulema	77	3,51E+10	üles_atma	4	6,92E+09	ära_selgima	4	3,14E+09
ära_lõppema	10	3,44E+10	üles_tõusma	4	6,92E+09	edasi_ minema	10	3,04E+09
ära_müüma	17	3,39E+10	maha_ sadama	5	6,89E+09	üles_panema	11	2,89E+09
välja_minema	41	3,36E+10	ära_katkuma	10	6,68E+09	ringi_keerama	4	2,87E+09
sisse_ajama	19	3,18E+10	kinni_katma	4	6,62E+09	kinni_hoidma	4	2,76E+09
peale_ hakkama	25	3,1E+10	välja_kiskuma	5	6,59E+09	üles_lendama	3	2,74E+09
välja_ noppima	8	2,95E+10	ära_leotama	5	6,18E+09	kokku_ vedama	6	2,71E+09

5. Kokkuvõte

Sõnadevahelise seose tugevuse mõõtmise statistikuid kasutatakse mitmetes arvutilingvistilistes rakendustes sõnadevahelise seose tugevuse määramiseks. Statistikute kasutamise eeliseks lihtsa sageduse kokkulugemise ees on asjaolu, et statistikud võtavad arvesse ka ühendit moodustavate sõnade eraldiesinemise sagedused. Lisaks sõnade eraldiesinemise sagedustele on statistikute väärtuste arvutamiseks vajalik leida nende koosesinemise sagedus, oodatav koosesinemise sagedus ning valimimaht. Sõnadevahelise seose tugevuse mõõtmisel võib lähtuda erinevatest alustest. Kindlas naabruses koosesinemisel arvestatakse vaadeldava sõna konteksti ehk akent, mille määrab uurija.

Käesolevas töös rakendasin Eesti murdematerjalil nelja statistikut: t-skoori, vastastikuse informatsiooni väärtust MI-d, hii-ruut statistikut ning log-tõepära funktsiooni (statistikute arvulised väärtused pole omavahel võrreldavad). Katsematerjaliks olid murdekorpuse morfoloogiliselt märgendatud ja lausestatud tekstid. Katsed tegin kolme murderühma peal eraldi.

Katsetulemuste pingeridade põhjal võib öelda, et omavahel sarnasemaid tulemusi annavad hii-ruut ning MI ja t-skoor ning log-tõepära funktsioon. Hii-ruut ja

MI eelistavad selgelt madala esinemissagedusega ühendeid, mille komponentide esinemissagedused korpuses on samuti madalad.

Võrreldes igas murderühmas nelja statistiku sadat kõrgeima väärtusega ühendit, võib näha, et ranniku- ja lõunaeesti murderühmas on väga sarnased ka hii-ruudu ning log-tõepära väärtused. Statistikutevahelised erinevused on väikseimad rannikumurde murderühmas ning suurimad põhjaeesti murderühmas. Üks põhjus on kindlasti materjali erinev hulk korpuses – põhjaeesti materjal on tunduvalt paremini esindatud. Omavahel kõige vähem kattuvad kõigis murderühmades MI ja t-skoori ning MI ja log-tõepära funktsiooni väärtused.

Ühtegi statistikut ei saa pidada teistest ühemõtteliselt paremaks.[6] Erinevad statistikud sobivad erinevat tüüpi ülesannete lahendamiseks. Statistiku sobivus sõltub kollokatiivse seose tüübist, korpuse suurusest, valdkonnast, keelest jmt. Näiteks oli antud katsete puhul üsna määrav lausestaja töö efektiivsus, sest mida korrektsemalt on osalaused eraldatud, seda adekvaatsem on kandidaatandmestik. Samuti võib statistikute töös näha erinevusi eri murderühmade vahel. Näiteks sobib MI paremini põhjaeesti murderühmale rakendades, sest annab madalad väärtused sagedastest sõnadest moodustunud mitte-ühendverbidele. Samuti toimib statistik ka lõunaeesti murderühmas, kuid siin esinesid sagedastest sõnadest moodustunud ühendid sageli ühendverbina, mis MI tööpõhimõttest tingituna saavad madalad väärtused.

Nii MI kui ka hii-ruut annavad kõrgemaid väärtusi madala sagedusega andmestikule, kuid sellise andmestiku jaoks paistab paremini töötavat siiski MI. Madala sagedusega ühendverbide eraldamiseks võiks MI-d kombineerida log-tõepära funktsiooniga. Log-tõepära funktsioon töötab selgesti kõige efektiivsemalt kõigis murderühmades, kuid MI eraldab väga hästi just madala esinemissagedusega ühendverbe. Seega ei oleks mõistlik valida ühte kindlat statistikut kogu materjali jaoks. Murrete ühendverbide leidmiseks sobib selgelt kõige paremini log-tõepära funktsioon, kuid parema tulemuse saamiseks peaks kasutama ka MI-d. Ranniku- ja lõunaeesti murderühmas võib madala esinemissagedusega ühendite leidmisel rakendada ka hii-ruut-statistikut.

Viidatud kirjandus

Eesti kirjakeele korpus. http://www.clut.ee/korpused/kasutajaliides/ (30.08.2009).

Eesti murrete korpus. www.murre.ut.ee (30.08.2009).

Evert, Stefan 2004. The Statistics of Word Cooccurences: Word Pairs and Collocations. Dissertation, Institut für maschinell Sprachverarbeitung, University of Stuttgart. http://elib.uni-stuttgart.de/opus/volltexte/2005/2371/pdf/Evert2005phd.pdf (30.08.2009).

Evert, Stefan 2008. Association Measures. http://www.collocations.de/AM/ (01.09.2009).

Evert, Stefan 2009. Corpora and collocations. – A. Lüdeling, M. Kytö (Eds.). Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, 1212–1249.

Stefanowitsch, Anatol; Gries, Stefan Th. 2009. Corpora and Grammar. – A. Lüdeling, M. Kytö (Eds.). Corpus Linguistics. An International Handbook. Berlin: Mouton de Gruyter, 933–952.

ISI – International Statistical Institute (The ISI Glossary of Statistical Terms). http://isi.cbs.nl/glossary/ (01.09.2009).

[6] Krenn ja Evert (2001) on muuhulgas väljendanud ka arvamust, et ükski meetod ei anna oluliselt paremaid tulemusi kui tavaline sageduse arvutamine (vt ka Uiboaed 2010).

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. Toulouse, France, 39–46.

Lindström, Liina; Müürisep, Kaili 2009. Parsing Corpus of Estonian Dialects. – Eckhard Bick, Kristin Hagen, Kaili Müürisep, Trond Trosterud (Eds.). Proceedings of the NODALIDA 2009 workshop. Constraint Grammar and Robust Parsing. May 14, 2009. NEALT Proceedings Series, 8. Odense: Northern European Association for Language Technology, 22–29.

Manning, Christopher D.; Schütze, Hinrich 2003. Foundations of Natural Language Processing. London: The MIT Press.

Quantum Computing and Quantum Information Processing. National Laboratory for Scientific Computing. http://virtual01.lncc.br/~giraldi/qc/itheory.html (01.09.2009).

Uiboaed, Kristel 2008. Ühendverbid Eesti murrete korpuses. Magistritöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut.

Uiboaed, Kristel 2010. Ühendverbid eesti murretes. – Keel ja Kirjandus, 1, 17–36.

Kristel Uiboaed (Tartu Ülikool). Uurimisvaldkonnad on korpuslingvistika, eesti murded, statistilised meetodid keeleteaduses.
kristel.uiboaed@ut.ee

STATISTICAL METHODS FOR PHRASAL VERB DETECTION IN ESTONIAN DIALECTS

Kristel Uiboaed
University of Tartu

The aim of this study was to assess different statistical methods of automatic collocations extraction from the corpus. To extract the collocations, association measures (AM) were applied and the association scores (AS) for the collocation candidates found in the corpus were calculated. An AS indicates the collocational strength between two words. An advantage of the AMs is the fact that in addition to the co-occurrence frequency, the marginal frequencies of collocating words are also taken into account. To calculate the AS, the following data is needed: co-occurrence frequency, marginal frequencies of collocating words, expected frequency and the sample size.

There are different approaches to applying AMs: words can be considered collocational only if they appear in the same collocational span, in one text unit (clause, sentence, utterance), or if they carry together some syntactic function. This paper attempts to apply AMs for phrasal verb detection from the Corpus of Estonian Dialects (CED). Texts of CED were morphologically tagged and parsed. Combinations of adverbs and verbs were extracted and AS was calculated for every collocation candidate. Experiments were run on three different dialect groups applying four different association scores: t-score, Mutual Information, chi-squared test and log-likelihood.

The results indicate that log-likelihood and t-score outperform MI and chi-squared test. The outcomes of different measures vary the most in the Northern dialect group. The best measure for dialect data in general is log-likelihood. However, MI and chi-squared test work well with low frequency data. In the Northern dialect group the best AM for low-frequency phrasal verb detection is MI, however, in the North-Eastern and Southern groups chi-square test works well for the same purpose. To achieve better results different scores should be combined.

Keywords: computational linguistics, corpus linguistics, dialectology, methods and tools, statistics, Estonian