Tartu Ülikool

Loodus- ja täppisteaduste valdkond

Matemaatika ja statistika instituut

Hanna Pook

Sageduste standardiseerimise mõju korrespondentsanalüüsi tulemustele eesti murretes esinevate elatiivi funktsioonide näitel

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendajad Mare Vähi ja Kristel Uiboaed

Tartu 2016

Sageduste standardiseerimise mõju korrespondentsanalüüsi tulemustele eesti murretes esinevate elatiivi funktsioonide näitel

Sageduste standardiseerimine on meetod võrdlemaks eri suurusega valimeid. Käesolevas bakalaureusetöös on uuritud, kas standardiseerimine avaldab korrespondentsanalüüsi tulemustele mõju. Töö aluseks on kümne eesti murde elatiivi funktsioonide sagedusandmestik. Standardiseerimata ja standardiseeritud andmete võrdlemiseks on rakendatud korrespondentsanalüüsi. Analüüsi käigus lükati ümber tööle püstitatud hüpotees ning järeldati, et standardiseerimine ei avalda korrespondentsanalüüsi tulemustele mõju.

Märksõnad: sageduste standardiseerimine, korrespondentsanalüüs, elatiiv, eesti murded

CERCS teaduseriala: statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika (P160)

The Effect of Frequency Normalisation on the Results of Correspondence Analysis as Shown on Elative Functions in Estonian Dialects

Frequency normalisation is a technique to compare different sized samples. The current study analyses whether frequency normalisation has an effect on the results of correspondence analysis. The studied data set consists of the frequencies of elative functions in ten Estonian dialects. Correspondence analysis has been applied to compare the raw and normalised data sets. In the course of the analysis the proposed hypothesis was rejected and it was concluded that frequency normalisation does not have an effect on the results of correspondence analysis.

Keywords: frequency normalisation, correspondence analysis (CA), elative, Estonian dialects

CERCS research specialisation: statistics, operation research, programming, actuarial mathematics (P160)

SISUKORD

SISSEJUHATUS

Käesolevas bakalaureusetöös on vaatluse all sageduse standardiseerimine ja selle mõju statistilise analüüsi tulemustele. Sageduse standardiseerimist kasutatakse näiteks keeleteaduses, psühholoogias ja sotsioloogias, kus on vaja võrrelda gruppe või tekste, mis ei ole võrdse mahuga. Selles töös on välja selgitatud, kas standardiseerimisel on korrespondentsanalüüsi tulemustele mõju või annavad standardiseeritud ja standardiseerimata andmed ühe nähtuse analüüsimisel sarnased tulemused.

Eesmärgiga ühendada oma peaeriala mastemaatiline statistika kõrvaleriala eesti ja soome-ugri keeleteadusega, on töö autor otsustanud uurida sageduste standardiseerimist just keeleteaduslikus kontekstis. Näiteks kasutatakse seda meetodit tihti murrete sagedusandmestike omavaheliseks võrdlemiseks, sest eri murrete materjali kättesaadavus varieerub laial skaalal ning seetõttu pole eri murrete valimid omavahel otseselt võrreldavad. Siinses töös on võrreldud elatiivi (seestütleva käände) funktsioonide kasutamist eesti murretes standardiseeritud ja standardiseerimata sageduste põhjal, kasutades selleks korrespondentsanalüüsi. Eelnimetatud analüüs on valitud seetõttu, et korrespondentsanalüüsi rakendatakse sageli keeleteaduslike andmete võrdlemiseks.

Peamine uurimisküsimus on välja selgitada, kas sageduste standardiseerimisel on mõju analüüsides saadud tulemustele. Hüpotees on, et sageduste standardiseerimine mõjutab korrespondentsanalüüsi tulemusi ja on seega objektiivse tulemuse saamiseks ning selle tõlgendamiseks vajalik. Seda hüpoteesi toetavad mitmed kirjanduslikud allikad (Adolphs 2006, Biber 1988, Biber jt 1998 jm), mis standardiseerimise kasutamist statistilistes analüüsides nõuavad. Täiendavalt on töös soovitud uurida, kas ja kuidas elatiivi funktsioonide kasutus eesti murretes üldse varieerub.

Töö on üles ehitatud järgnevalt. Esimeses peatükis antakse ülevaade tekstiandmete võrdlemisest ja sageduste standardiseerimisest. Teine peatükk kirjeldab elatiivi ajalugu ning eri autorite käsitluste põhjal selle funktsioone. Kolmandas peatükis on tutvustatud töös kasutatud andmestikku, selle tausta ning kodeerimisprotsessi ja -põhimõtteid. Neljas peatükk selgitab korrespondentsanalüüsi – meetodit, mida on standardiseeritud ja standardiseerimata andmestike võrdlemiseks kasutatud. Viiendas peatükis on esitatud kvantitatiivse analüüsi tulemused ning järeldused.

Töö kirjutamiseks on kasutatud tekstitöötlustarkvara Microsoft Word 2010. Analüüsi läbiviimiseks on kasutatud statistikapaketti SAS 9.4 ja tabelarvutustarkvara Microsoft Excel 2010.

Autor tänab juhendajat Mare Vähi väärtuslike nõuannete ja matemaatiliste lahenduste eest ning juhendajat Kristel Uiboaeda töö idee, täienduste ja parandusettepanekute eest.

1. SAGEDUSTE STANDARDISEERIMINE

Keeleteaduses, aga ka paljudes sotsiaalteadustes, on tavaline olukord, kus analüüsitavate andmete hulk on piiratud ja lisaandmete kogumine on erinevatel põhjustel (ajaloolised, eetilised, praktilised vm) kas raskendatud või üldse võimatu. Selline olukord tekitab probleeme andmete võrdlemisel, sest valimite mahud varieeruvad tihti laial skaalal ning pole seetõttu sellisel kujul omavahel võrreldavad.

Niisamuti peab ka korpustel[1] põhinevates analüüsides, mis käsitlevad keelenähtuste sagedusi üle erinevate tekstide või alamkorpuste, jälgima, et analüüsitavad sagedused oleksid võrreldavad. Kui uurida sama nähtust kahes erinevas tekstis, millest ühes on kaks korda rohkem sõnu kui teises, võib pikemas tekstis uuritavaid nähtuseid ka kaks korda rohkem esineda. Seega ei osuta suured lihtsagedused alati keelenähtuse sagedamale kasutusele valitud alamkorpuses ning neid ei saa tekstide omavaheliseks võrdlemiseks kasutada (Biber 1988: 14).

Üks võimalus keelenähtuse sageduse korrektseks võrdlemiseks eri tekstides on kasutada selle suhtelist sagedust, mis näitab nähtuse esinemise protsenti tekstis. Seda meetodit kasutatakse sageli suurte tekstikorpuste analüüsimisel (Adolphs 2006: 43). Suhteline sagedus arvutatakse valemiga

$$f_i = \frac{n_i}{N},$$

kus n_i on i-nda keelenähtuse esinemissagedus ja N on sõnade arv kogu tekstis.

Suhtelise sageduse kasutamise probleem seisneb selles, et iga statistilise meetodi sisendiks protsendid ei sobi. Ka selles töös kasutatav korrespondentsanalüüs nõuab

[1] Tekstikorpus on suulistest või kirjalikest tekstidest koosnev elektrooniline andmekogu. Üldjuhul on korpused ka märgendatud (kõikidele sõnadele on lisatud nt vormi- või muu lingvistiline info) ja varustatud metaandmetega.

sisendiks just sagedusandmestikku, järelikult kõikidel juhtudel suhteliste sageduste arvutamisest tekstide või korpuste võrdlemiseks ei piisa.

Alternatiiv suhtelise sageduse arvutamisele on kasutada sageduste standardiseerimist, mis ühtlustab lihtsagedusi selliselt, et neid oleks võimalik omavahel võrrelda. Standardiseeritud sagedus näitab, mitu korda uuritav nähtus esineb baasiks valitud arvu sõnade hulgas (Biber jt 1998: 33) ja see arvutatakse järgneva valemiga (Biber 1988: 14):

$$sdn_{ij} = \frac{n_{ij} \cdot baas}{N_j},$$

kus n_{ij} on *i*-nda keelenähtuse lihtsagedus *j*-ndas tekstis või alamkorpuses ja N_j on kogu sõnade arv selles *j*-ndas tekstis. Standardiseerimisbaas tuleks valida vastavalt alamkorpuse suurusele: kui võrreldavad tekstid on kõik umbes 1000 sõna pikkused, siis peaks ka baas olema 1000 (Biber 1998: 264). Baasiks võib lisaks ligikaudsele numbrile valida ka korpuse keskmise sõnade arvu, mida on kasutanud näiteks Uiboaed (2013).

Oluline on märkida, et kui nähtuse lihtsagedus on väga väike, siis võib selle standardiseerimine saadud tulemusi moonutada (Adolphs 2006: 43). Samuti kui valida liiga suur baas (näiteks baas 1000 tekstidele, milles on vähem kui 100 sõna), võivad standardiseeritud sagedused kunstlikult suureneda (Biber 1998: 264).

2. ELATIIV

Eesti keeles on sisekohakäänded illatiiv ehk sisseütlev, inessiiv ehk seesütlev ja elatiiv ehk seestütlev, väliskohakäänded on allatiiv ehk alaleütlev, adessiiv ehk alalütlev ja ablatiiv ehk alaltütlev. Selline tänapäevane kohakäänete süsteem kujunes välja juba läänemeresoome algkeeles (Rätsep 1979: 27). Järgnevates alapeatükkides on lühidalt tutvustatud elatiivi (ja teiste kohakäänete) ajalugu ning on antud ülevaade eri autorite elatiivi funktsioonide käsitlustest.

2.1. Elatiivi ajalooline taust

Enne seda, kui läänemeresoome algkeeles eristusid sise- ja väliskohakäänded, esines Uurali algkeeles küll eesti keelele omane kohakäänete kolmeaspektilisus, aga kolmeks kohakäändeks selles on Rätsepa (1979: 27) järgi peetud *k*-latiivi, *na*-lokatiivi ja *ta*-separatiivi. *na*-lokatiivil ja *ta*-separatiivil ei ole tänapäeval enam kohta näitavat funktsiooni (esimesest on saanud olev, teisest osastav kääne), *k*-latiiv on täiesti kasutuselt kadunud.

Nii sise- kui ka väliskohakäänete lõpud on liitkäändelõpud, koosnedes kahest käändeelemendist: sisekohakäänetel on ühine element *-s*, millele on liitunud vanad kohakäändelõpud, väliskohakäänetel on ühiseks elemendiks *-l*. *-s* oli vana läänemeresoome ja volga keeltele ühine latiivne käändelõpp, mis on eesti keeles lõpuna säilinud veel üksikutes adverbides, nt *alaspidi*, *siis*, *taas* (Rätsep 1979: 27–28).

Praeguses kirjakeeles on elatiivi lõpuks *-st*, aga varasemalt oli käändelõpuks *-sta*/*-stä*, mis koosnes latiivilõpust *-s* ja sellele liitunud separatiivilõpust *-ta*/*-tä*. 13. sajandil, kui eesti keeles toimus lõpukadu, kadus ka vokaal elatiivi lõpust. Lisaks noomenitele esineb elatiivi lõpp ka *ma*-infinitiivi puhul (Rätsep 1979: 50).

17. sajandil, kui hakati eesti keele grammatikaid kirjutama, lähtuti nendes kas ladina või saksa keelest. Kuna aga eesti keele käändeparadigma erineb nende keelte käänetest suuresti, oli grammatikutel raske eesti keele käändesüsteemi adekvaatselt kirjeldada. Seega näiteks esimeses eesti keele grammatikas, mille kirjutas Heinrich Stahl 1637. aastal, on sisekohakäänetest vormiliselt esindatud ainult ainsuse seestütlev (*öhest Jummalast*), aga selle vormid on paigutatud nii ablatiivi ja genitiivi alla (Ross 1997: 185). 1648. aastal avaldatud Gutslaffi grammatikas esitatakse kohakäänetest ainult seestütleva vorm genitiivina ja alaleütleva vorm daativina. 1660. aasta Gösekeni grammatikas esitatud käändesüsteem sarnaneb Stahli omaga ning 19. sajandi alguses avaldatud Hupeli grammatika teine trükk esitab sisekohakäänetest vaid seestütleva vormi, aga juba ablatiivi tähenduses ja paralleelselt alaleütlevaga. Ainsaks erandiks varasemate grammatikute seas on Hornung, (grammatika ilmus 1693. aastal), kes küll esitab samuti alalt- ja seestütleva vormi ablatiivi paralleelvõimalustena, kuid lisaks neile toob ta ablatiivi variantidena välja ka sisseütleva lõpu *-sse* ja seesütleva lõpu *-s* (Ross 1997: 186–187).

Grammatikates sai elatiiv omaette käändeks alles 19. sajandil. Esialgu liigitas Otto Wilhelm Masing elatiivivormid lihtsalt *s*-käänete hulka, aga Ahrensi 1853. aastal avaldatud „Eesti keele Tallinna murde grammatikas“ sai see esmakordselt nimetuse elatiiv. Eestikeelse nime – seestütlev – andis käändele 1880. aastal Karl August Hermann (Rätsep 1979: 50).

Ka murretes on elatiivi käändelõpuks olnud tavaliselt *-st*. Ainsaks erandiks on Kodavere murrak, kus *st* on muutunud *ss*-iks (sõna lõpul *-s̀*) ja seega elatiivilõpp on selles murrakus *-s̀*. Seetõttu langeb elatiivivorm seal kokku translatiiviga, kuna ka *ks* on muutunud *ss*-iks (Rätsep 1979: 51).

2.2. Elatiivi funktsioonid

Kuigi prototüüpselt on elatiivil kohatähendus, on sellel käändel hulgaliselt muidki funktsioone. Esmalt on esitatud elatiivi funktsioonide kirjeldused esimestes

grammatikates, mis hakkasid eesti keele grammatikat seletama soome keele eeskujul. Kuigi nendes loetletud elatiivi funktsioonid erinevad mõnevõrra tänapäevastest käsitlustest, on need varasemad kirjeldused välja toodud just seetõttu, et edaspidistes peatükkides kasutatud murdetekstide lindistused on kogutud peamiselt inimestelt, kes on sündinud 19. sajandi teisel poolel või 20. sajandi alguses. Seega pole uurimismaterjal päris tänapäevase keelekasutusega.

Ahrens (2003 [1853]: 359–365) on oma grammatikas loetlenud, et elatiiv märgib

1. liikumist millestki välja, nt *wöta raud tulest, sibulad wötavad vee silmist wälja*;
2. seisundit, milles (millest lähtudes) midagi toimub, nt *joobnust peast hakkas riidlema, rukis tahab hädast leigata*;
3. ajahetke, millest alates aega arvestatakse, nt *sest pääwast hakkas pödema, maast madalast, sügisest talwe*;
4. ainet, millest asi koosneb, nt *puust ja rauast, tema on suurest soost sündinud*;
5. põhjust, millest tagajärg tuleneb, nt *see tuleb kadedusest, tuluke hakkas ühest sädemest öhkuma, ega temale ole Jumalast seda tarkust antud*;
6. asja, millest on osa ära võetud või mille kohta on midagi mõeldud või öeldud, nt *ta söi mu leiwast ja jöi mu piimast, köige wanem neist wendadest*;
7. isiku või asja osa, mille kohta öeldu käib, nt *tema lambad on muist mustad, arust lühikene, oleksin ma seda hingest teadnud, mul on weel üks Annest tütar* (õdesid-vendi peetakse üheks tervikuks ja iga last selle terviku üheks osaks);
8. seda, millest asi eraldatakse või kaugele jääb, nt *tahab naesest äralahutada, läks meilt ära, sinna jäi minust wanu puid maha, pühi laud tolmust puhtaks;*
9. seda, millest ollakse ilma või mis puudub, nt *nemad on kirjatundmisest ilma, nahk oli karwast paljas, töötegijaist oli neil puudu*;
10. asja, mille kohta käib sõnade *läbi*, *mööda*, *sisse*, *välja*, *üles*, *alla*, *üle* juurde liituv verb, nt *läks jala jöest läbi, tulin heinamaast läbi, jooksis uksest wälja*;
11. kaupa, mille eest makstakse mingit hinda, nt *maksin kübarast neli rubla, mis herra wakast peab, andis 80 rubla hobusest*;
12. keskvõrde korral seda, millega asja võrreldakse, nt *ta on minust suurem, kül sa jääd minust weel waesemaks*.

Wiedemanni (2011 [1875]: 365–368) „Eesti keele grammatikas“ esitatud elatiivi funktsioonid kattuvad suuremalt jaolt Ahrensi omadega. Tema toob välja järgmised elatiivi kasutusvõimalused.

1. Elatiivi kasutatakse mitmesuguste väljendite puhul, mis märgivad lahkumist või eemaldumist, nt *tuli sealt külast, sain sest kimbust lahti, taganes minust.*
2. Elatiiv väljendab olukorda, milles või millest lähtudes midagi tehakse, nt *rākis unest, kainest peast, lahkest südamest jumalat paluma.*
3. Elatiiv märgib suhet „millest alates“ ajas ja ruumis, nt *lapsepõlwest sādik, māst madalast, hommikust õhtuni.*
4. Elatiiv märgib, millest miski koosneb või kust on pärit, nt *seda tehakse rauast, sinust ei sā inimest, karu-nahast kazukas.*
5. Elatiiv märgib millegi ajendit, põhjust, nt *se tuleb kadeduzest, wihmast märg.*
6. Elatiiv märgib osa mingist tervikust, nt *jõi minu õllest, kõige nōrem sinu wendadest, mis sa minust naerad.*
7. Elatiivi kasutatakse ladina *de*-eessõnalise ablatiivi puhul tegusõnade *rääkima, vaikima, kuulma, mõtlema* jne järel, nt *sest aśjast olen ju kūlnud, rā̄gib palju omast tegudest, mis sa sest ütled.*
8. Elatiivi kasutatakse lähemaks määratlemiseks, osutades selle abil mingile osale või erijoonele, mille kohta öeldu eriti käib, nt *õńsad nēd, kes puhtad südamest on; läheb näust siniseks*, lisaks kasutatakse seda liikumist väljendavate tegusõnade puhul, et osutada kohta mingi adverbiga seoses, nt *läks uksest wälja, tuli mäest alla, hüpas mulgust üle.*
9. Elatiiviga väljendatakse seda, mille eest makstakse mingit hinda, nt *sai kolm rubla wakast, mis sa sest wańkrist maksid.*
10. Elatiivi kasutatakse võrdluste puhul, nt *ta on minust sūrem, küll ta sāb wennast wēl rikkamaks, tūl on eilsest tagasi.*
11. Teatud territooriumitel kasutatakse elatiivi ajaväljendites adessiivi ja faktiivi asemel, vastusena küsimustele „millal“, „kui kaua“, nt *õigest ajast* ‘õigeks ajaks’, *wēl mõnest ajast* ‘veel mõneks ajaks’.

12. Elatiivi kasutatakse veel ühel väga omapärasel ja raskesti mõistetaval viisil, nt *üks Jānist laps* ‘laps nimega Jaan’, *Reinust poeg on tema jūres, mull on üks Maŕdist wend.*

Hilisematest elatiivi funktsioonide kirjeldustest on välja toodud „Eesti keele grammatikas“ ja „Eesti keele käsiraamatus“ esitatu. EKG (1995: 57–58) kohaselt on elatiivi ülesandeks väljendada:

1. separatiivse tähendusega kohta, nt *leidsin sahtlist märkmiku, saabusime pulmast*;
2. separatiivse tähendusega aega, nt *nii on olnud juba vanast ajast*;
3. separatiivse tähendusega seisundit, nt *poiss toibus minestusest*;
4. materjali, millest miski on (tehtud) või koosneb, nt *paneelidest majad, seltskond koosnes erinevatest inimestest*;
5. tervikut, eriti võrdlusalusena komparatiivse adjektiivi või adverbi juures, nt *üks rääkis teisest tükk maad rohkem, Jüri oli poistest kõige nutikam*;
6. põhjust, nt *sõrmed olid külmast kanged*;
7. asja osa, mida verbi tegevus haarab või mida adjektiiv iseloomustab, nt *näost kaame, peast haavata saanud*;
8. (infiniittarindi) tegevussubjekti, nt *mees oli murest murtud, justkui nagu kurjast vaimust vaevatud*;
9. asja, millega puudub või katkeb ühendus, nt *loobusin sellest tööst, kõik nagu hoiduksid sellest paigast eemale*;
10. teatud kindlate verbide rektsioonilise laiendina ilma selge tähenduseta, nt *ta peab sellest luuletajast väga lugu, mõtlen sageli sellest raamatust.*

EKK (2007: 248–249) järgi saab elatiiviga väljendada

1. lähtekohta, nt *tulin linnast*;
2. algusaega, nt *ootasin hommikust õhtuni*;
3. lähteseisundit, nt *toibus minestusest*;
4. võrdlusalust, nt *koer oli kassist suurem*;
5. materjali, nt *laud oli puidust*;
6. kellest või millest midagi tuleb või saab, nt *sinust saab arst*;

7. põhjust, nt *käed külmast kanged*;
8. passiivse infiniittarindi tegevussubjekti, nt *murest murtud*;
9. lisandit, nt *üliõpilasest õde*;
10. objekti, millele tegevus on suunatud, nt *ta ei saanud sellest aru, miks sa ei rääkinud mulle oma murest, mis sa temast kiusad*;
11. midagi, millega puudub või katkeb ühendus, nt *ta jäi palgast ilma, loobus oma ideest*;
12. on selge tähenduseta, *nt vaimustub kõigest väga kergesti, kõik oleneb olukorrast*;
13. rühma või tervikut, mille osaks miski on, nt *Jüri oli poistest kõige nutikam, ta on näost kaame, kolm inimest sellest seltskonnast olid mulle võõrad.*

Niisiis võib seestütleval käändel olla hulgaliselt erinevaid funktsioone, kusjuures ükski neist loeteludest ei ole tegelikult ammendav. Nende kirjelduste alus on valdavalt kirjakeel ja mitte ebastandardsemad keelekujud nagu seda on näiteks murdematerjal. Murrete kohta selliseid põhjalikke elatiivi funktsioonide kirjeldusi ei ole koostatud, aga võib eeldada, et suuremas osas kattuvad need kirjakeele funktsioonidega.

3. ANDMESTIK

3.1. Eesti murrete korpus

Töö aluseks olev uurimismaterjal on saadud eesti murrete korpusest. Eesti murrete korpus on elektrooniline andmete kogu, mis sisaldab autentseid tekste kõikides eesti murretes. Korpuse materjal koosneb võimalikult vanapärastest murdetekstidest, millest suurem osa on kogutud 1960.–1970. aastatel. Kõikidest tekstidest on olemas ka helisalvestised, mida säilitatakse Eesti Keele Instituudis (EKI) ning Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivis. Korpust on alates 1998. aastast koostatud EKI ning Tartu Ülikooli eesti ja üldkeeleteaduse instituudi koostöös (EMK 2015).

Eesti murrete korpus koosneb järgmistest osadest:

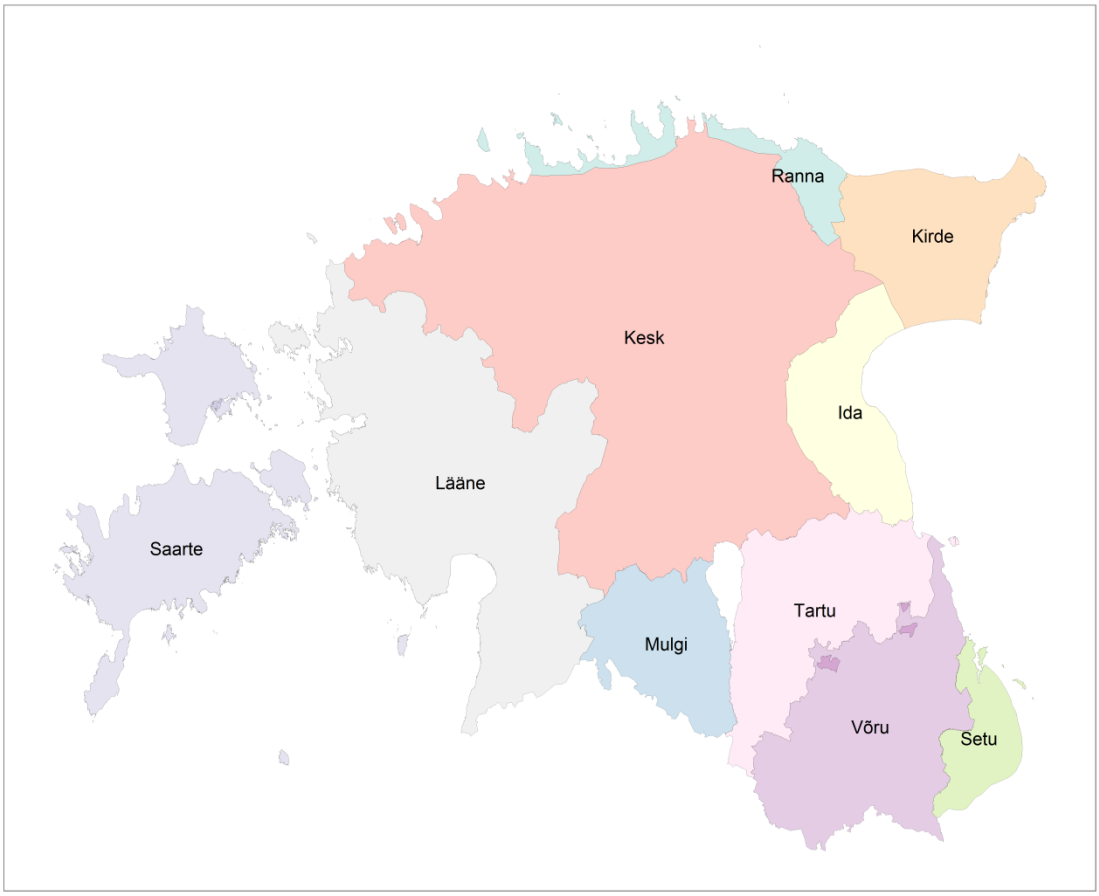
1. helisalvestised,
2. foneetilises transkriptsioonis[2] murdetekstid,
3. lihtsustatud transkriptsioonis murdetekstid,
4. morfoloogiliselt märgendatud tekstid,
5. andmed keelejuhtide[3], lindistuste ja litereeringute kohta (EMK 2015).

Käesolevas töös on kasutatud morfoloogiliselt märgendatud tekste, mis on XML-formaadis ja kus igale märksõnale tekstis on lisatud lemma (sõna algvorm kirjakeeles), sõnaliik, sõnavorm ja fraas. Need tekstid on vabalt kättesaadavad murdekorpuse otsingumootori kaudu <http://www.murre.ut.ee/mkweb/> (EMK 2015). Murdekorpust täiendatakse pidevalt, seega viidatud otsingumootor tagastab rohkem märksõnu, kui on kasutatud siinse töö andmestikus, mis on kogutud 01.10.2014.

Korpuses on eesti murded jagatud kümneks: põhjaeesti murderühma alla kuuluvad keskmurre, läänemurre, saarte murre ja idamurre, lõunaeesti murderühma alla kuuluvad

[2] Foneetiline transkriptsioon on kõne hääldusliku võimalikult täpse ülesmärkimise süsteem.
[3] Keelejuht on keelt või murret usaldusväärselt valdav inimene, kellelt uurija keeleainestikku kogub.

Tartu, Võru, Mulgi ja Seto murre ning kirderanniku murderühma alla kuuluvad kirdemurre (ehk Alutaguse murre) ja rannamurre (EMK 2015). Seega erineb korpuse murdejaotus traditsioonilisest jaotusest, mis murdeõpiku „Eesti murded ja kohanimed“ kohaselt ei erista Võru ja Seto murdeid, vaid peab Seto murret Võru murde murrakurühmaks (Pajusalu jt 2009: 55). Kõik murded on omakorda kihelkonniti jaotatud murrakuteks (EMK 2015). Eesti murrete kaart on esitatud joonisel 1.



Joonis 1. Eesti murrete korpuse murdejaotus. Põhjaeesti murderühm: ida-, kesk-, lääne- ja saarte murre; lõunaeesti murderühm: Mulgi, Seto, Tartu ja Võru murre; kirderanniku murderühm: kirde- ja rannamurre (kaart kohandatud EKI 2014 põhjal).

3.2. Andmete kirjeldus ja kodeering

Murdekorpuse tekstid pärinevad 267-lt keelejuhilt, kellest üle kahe kolmandiku on naised. Keelejuhid on sündinud vahemikus 1865–1919, 60% neist vahemikus 1877–1892. Töös kasutatud murdetekstides on kokku 834 311 morfoloogiliselt märgendatud sõna. Nende jagunemine murrete lõikes on esitatud tabelis 1. Kõige paremini on esindatud saarte-, lääne- ja keskmurre ning kõige vähem sõnu on kogutud kirde-, ida- ja Seto murdest.

Tabel 1. Murdekorpuse sõnade arv murrete lõikes (seisuga 01.10.2014).

murre	**sõnade arv**
saarte	166 898
lääne	154 400
kesk	130 086
Võru	70 038
Tartu	65 591
Mulgi	63 516
ranna	51 667
kirde	47 660
ida	45 280
Seto	39 175
kokku	**834 311**

Korpuse otsingumootoris elatiivi järgi otsides on koostatud andmestik, mis koosneb 7129-st seestütlevas käändes märksõnast. Iga märksõnaga koos on esitatud sellele eelnev ja järgnev kontekst (kuni 10 sõna), lemma (sõna algvorm kirjakeeles), sõnaliik ja informatsioon keelejuhi kohta (murre, murrak ja küla). Oluline on mainida, et murdekorpus on käsitsi märgendatud, seega andmetes võib sõnaliigi, vormi, lemma vm märkimisel esineda vigu või ebaühtlusi.

Tabelis 2 on esitatud elatiivis märksõnade jagunemine murdeti. On näha, et seestütlevas käändes märksõnade sagedused ei ole üheses vastavuses kõikide murdetekstide sõnade sagedustega, seega on alust eeldada, et elatiivi (või mõne kindla elatiivi funktsiooni) kasutus eri murrete vahel varieerub.

Tabel 2. Elatiivis märksõnade jagunemine murrete lõikes.

murre	**sõnade arv**
saarte	1560
lääne	1332
kesk	1071
ranna	592
Mulgi	539
Võru	469
ida	461
kirde	425
Tartu	413
Seto	267
kokku	**7129**

Kuna murdekorpus on märgendatud ainult morfoloogiliselt, siis elatiivi funktsioone automaatselt otsida pole võimalik. Seetõttu on töö autor igale märksõnale vastava funktsiooni ise lisanud. Funktsioonide kodeering põhineb peamiselt EKG ja EKK elatiivi funktsioonide kirjeldusel ning on järgmine (sulgudes on lisatud tabelites ning analüüsis kasutatud funktsioonide lühendid):

- abstraktne lähtekoht (absk),
- algusaeg (aeg),
- lähtekoht (koht),

- materjal (mat),
- põhjus (põh),
- püsiühend[4] (pü),
- rektsioon[5] (rek),
- lähtekoht sündmuselt või organisatsioonist (sündk),
- rühm või tervik, mille osaks miski on (ter),
- resultatiiv ehk tulemus (tul),
- võrdlusalus (vral).

Lisaks nendele funktsioonidele on andmestikus eraldi kodeeritud ka fraaside (f) esimene osa (nt ***nuorest*** *piast* (RAN), ***linatsõst*** *rõivast* (VÕR), ***uijest*** *teest* (LÄÄ), aga ka ***mõlemist*** *poolt* (KES)), vead (e) murdekorpuse andmetes (peamiselt juhud, kus märksõnale on lisatud morfoloogilise märgendina elatiiv kui märksõna tegelikult elatiivis ei ole) ja muud märksõnad (muu), mille kasutus ei sobitunud ühegi eelnevalt mainitud funktsiooni alla või mille funktsiooni ei olnud võimalik määrata (näiteks pooleli jäänud lausete puhul).

Kodeerimise tulemusel saadud märksõnade funktsioonide esinemissagedused murrete lõikes on esitatud tabelis 3.

[4] Püsiühend on keeles laialt käibel olev tavapärane sõnade ühend, millele on omane osade tähenduslik kokkukuulumine.
[5] Rektsioon ehk sõltumus on sõnadevaheline seos, kus põhjaks oleva sõna tähendus määrab ära tema laiendiks oleva sõna käände- või pöördevormi või temaga seotud kaassõna (Mäearu 2011: 3).

Tabel 3. Elatiivi funktsioonide esinemissagedused murrete lõikes.

	ida	kesk	kirde	Mulgi	ranna	saarte	Tartu	lääne	Seto	Võru	kokku
absk	9	30	2	8	5	20	13	7	1	13	108
aeg	20	57	20	22	17	61	11	62	9	13	292
koht	172	402	165	216	280	641	165	514	100	170	2825
mat	63	214	62	75	76	343	80	317	62	83	1375
põh	4	31	27	9	8	21	1	22	1	6	130
pü	33	28	7	13	13	30	18	32	11	9	194
rek	32	114	34	62	45	133	52	134	32	61	699
sündk	5	12	9	6	7	20	8	10	3	2	82
ter	10	8	6	2	3	4	0	6	0	2	41
tul	9	5	0	6	5	3	0	10	1	6	45
vral	12	13	14	11	7	17	6	18	1	6	105
muu	9	23	15	30	54	103	19	34	19	43	349
f	57	99	60	75	64	154	39	154	23	52	777
e	26	35	4	4	8	10	1	12	4	3	107
kokku	461	1071	425	539	592	1560	413	1332	267	469	7129

4. METOODIKA: KORRESPONDENTSANALÜÜS

Korrespondentsanalüüs on kvalitatiivsete tunnustega andmete analüüsimise meetod. Kõige lihtsamal kujul on sellega võimalik analüüsida kahemõõtmelist risttabelit, mille tulemusena saadakse numbrilised väärtused nii rea- kui ka veerutunnusele. Saadud väärtused peaksid kirjeldama võimalikult palju kahe tunnuse vahelisest seosest. Tavaliselt kujutatakse neid tunnuseid kahemõõtmelisel graafikul, millele on paigutatud saadud väärtuste paarid, mis võimaldab lihtsalt visualiseerida rea- ja veerutunnuse sarnasusi ning erinevusi (Pärna 1993: 3–4). Korrespondentsanalüüsi on võimalik rakendada erinevates programmides nagu näiteks SAS, R, SPSS, BMDP jm. Antud analüüsi kasutatakse sageli ka keeleteaduslike andmete võrdlemiseks, mistõttu ongi sageduste standardiseerimise mõju korrespondentsanalüüsi põhjal uuritud.

Olgu *n* erinevat vaatlust, mis on jagatud kahe tunnuse *A* ja *B* järgi. Olgu tunnusel *A* kokku I väärtust $(A_1, A_2, \dots, A_I)$ ja tunnusel *B* kokku J väärtust $(B_1, B_2, \dots, B_J)$, mis moodustavad risttabeli $N: I \times J = (n_{ij})$. Antud maatriksi rea-, veeru- ja kogusummad avalduvad vastavalt valemitega (Pärna 1993: 5):

$$n_{i+} = \sum_j n_{ij},$$

$$n_{+j} = \sum_i n_{ij},$$

$$n = \sum_i \sum_j n_{ij}.$$

Järgmisena defineeritakse suhtelised sagedused risttabelis *N* (Pärna 1993: 6):

$$f_{ij} = \frac{n_{ij}}{n},$$

$$f_i = \frac{n_{i+}}{n},$$

$$f_j = \frac{n_{+j}}{n},$$

$$f_j^i = \frac{f_{ij}}{f_i} = \frac{n_{ij}}{n_{i+}},$$

$$f_i^j = \frac{f_{ij}}{f_j} = \frac{n_{ij}}{n_{+j}}.$$

Maatriksi reaprofiil *i*-nda rea jaoks on vektor$f_B^i = (f_1^i, f_2^i, \dots, f_J^i)^T$ ja veeruprofiil *j*-inda veeru jaoks on vektor $f_A^j = (f_1^j, f_2^j, \dots, f_I^j)^T$ (Pärna 1993: 6). Rea- ja veeruprofiilid defineerivad kaks punktipilve vastavalt *J*- ja *I*-mõõtmelises eukleidilises ruumis (Greenacre 1984: 85). Reaprofiilide pilv avaldub valemiga

$$N_B(A) = \{f_B^i | i = 1, \dots, I\}$$

ja veeruprofiilide pilv valemiga

$$N_A(B) = \{f_A^j | j = 1, \dots, J\}.$$

Korrespondentsanalüüsis on igal profiilil oma punktipilves kaal, milleks on *i*-nda rea korral marginaaltõenäosus f_i (või f_j *j*-nda veeru korral). Seega on pilv kaalutud punktide kogu ja pilve keskpunkti saab defineerida kui selle masskeset ehk kõikide pilve elementide kaalutud keskmist (Pärna 1993: 6).

Eksisteerigu profiilid mitmemõõtmelises ruumis ja olgu tarvis leida madala dimensiooniga alamruum, mis oleks võimalikult lähedal kõikidele pilve punktidele. Greenacre (1993: 45–46) järgi ongi sellise alamruumi leidmine korrespondentsanalüüsi eesmärk. Iga profiilipunkti jaoks pilves leitakse χ^2-kaugus punkti ja pilve keskpunkti vahel, mis arvutatakse valemiga (Pärna 1993: 8)

$$d_i^2 = \sum_j \frac{\left(f_j^i - f_j\right)^2}{f_j}.$$

Profiili lähedus alamruumile leitakse valemiga

$$\sum_i f_i d_i^2,$$

kus f_i on *i*-nda profiili kaal (Greenacre 1993: 46).

Profiilide varieerumist pilve keskpunkti suhtes näitab inerts, mis on pilves $N_B(A)$ defineeritud kui *I* profiilipunkti kaalutud keskmine kaugus pilve keskpunktist (Pärna 1993: 9):

$$in(A) = \sum_i f_i d_i^2.$$

5. ANALÜÜS JA TULEMUSED

Selleks, et standardiseerimise mõju uurida, on töö autor koostanud kolm andmestikku. Esimeses neist (esitatud peatükis 3.2 tabelis 3) on elatiivi funktsioonide lihtsagedused. Järgmises kahes on sagedused standardiseeritud. Nagu varasemalt mainitud, saab standardiseeritud sagedusi leida valemiga

$$sdn_{ij} = \frac{n_{ij} * baas}{N_j},$$

kus antud andmestiku puhul on n_{ij} i-nda elatiivi funktsiooni lihtsagedus j-ndas murdes ja N_j kõikide sõnade arv j-ndas murdes. Standardiseerimisbaasid on valitud murdetekstide sõnade arvu põhjal. Esimene baas on korpuse keskmine sõnade arv, milleks on 83 431 sõna. Teiseks baasiks on vastavalt kirjanduslike allikate soovitusele valitud 100 000, mis on suuruselt lähedane murdekorpuse alamkorpuste (ehk murrete) mahtudele. Kaks baasi on valitud selleks, et näha, kas neil on erinev mõju tulemustele või mingi selge erinevus võrreldes standardiseerimata andmetega.

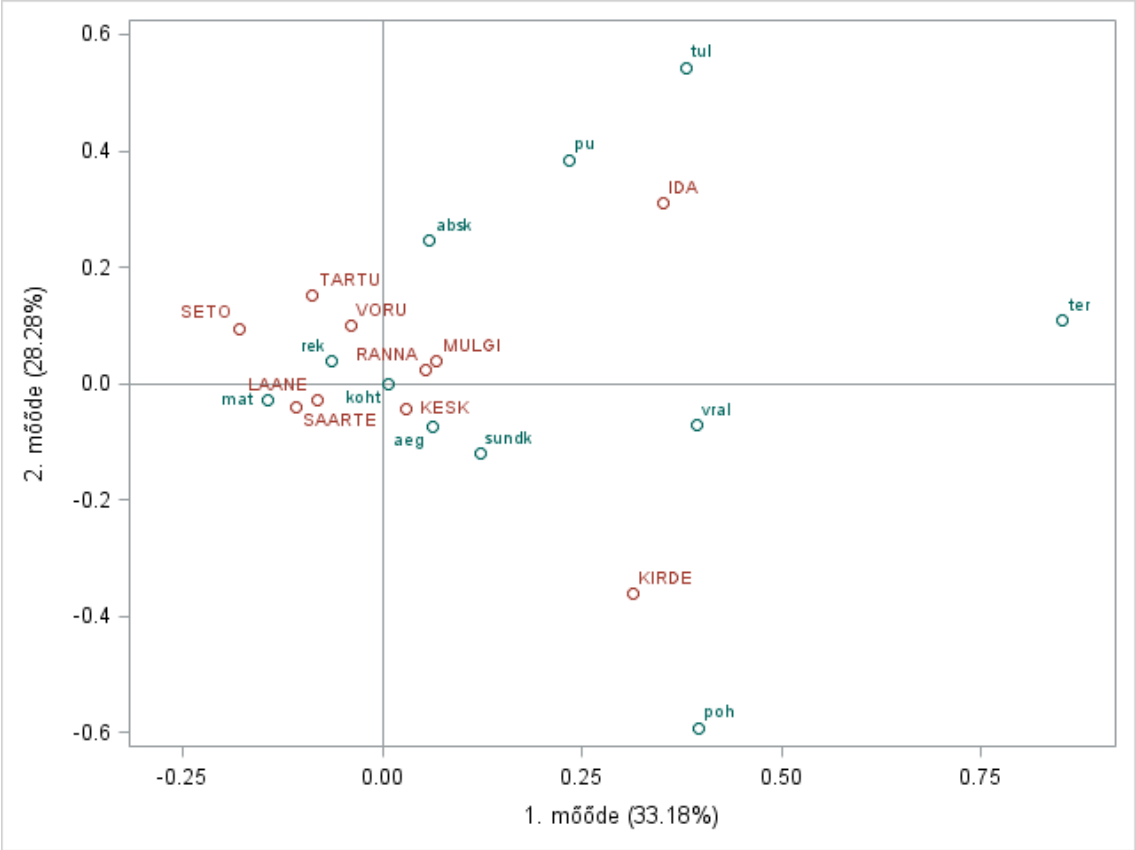
Kõikidest andmestikest on enne korrespondentsanalüüsi rakendamist jäetud välja fraasi osad, kodeerimisvead ja muud funktsioonid, mille analüüsimine ei annaks sisulisi tulemusi.

Järgmised alapeatükid esitavad analüüsi tulemused kolme andmestiku põhjal.

5.1. Standardiseerimata andmed

Statistikapaketiga SAS on standardiseerimata andmestiku põhjal tehtud korrespondentsanalüüs, mille tulemused on esitatud joonisel 2. On näha, et esimene mõõde kirjeldab 33% ja teine mõõde 28% koguinertsist, seega kokku üle 60% (vt ka

lisa 1). Graafiku keskel paiknevad telgjooned tähistavad mõlema mõõtme keskmist profiili.



Joonis 2. Korrespondentsanalüüsi graafik standardiseerimata andmete põhjal.

Nii esimese kui ka teise mõõtme põhjal eristub üks suurem grupp, mis paikneb pilve keskpunktis ja mis sisaldab kõiki murdeid peale ida- ja kirdemurde. Vertikaalse mõõtme põhjal saab seda gruppi jagada veel kaheks osaks, millest esimesse kuuluvad Tartu, Seto ja Võru murre ning teise ranna-, Mulgi-, saarte, lääne- ja keskmurre. Selline paigutus graafikul annab alust eeldusele, et elatiivi kasutatakse murretes pigem sarnastes funktsioonides ja sarnase sagedusastmega.

Prototüüpne kohatähendus (koht) on ootuspäraselt graafiku keskpunktis (näide 1) ja murretevahelised erinevused on seega tingitud pigem teistest elatiivi funktsioonidest. Samuti on sagedaselt kasutatavad funktsioonid algusaeg (aeg), materjal (mat) ja

rektsioon (rek) (näited 2–4), mida kinnitab ka graafik, kus need funktsioonid asuvad keskpunkti ümbruses. See tähendab, et neid funktsioone kasutatakse ühtlaselt kõikides murretes.

(1) siit **naabridalust** varastatti ükskord kaks obust ära (LÄÄ)
(2) kellä **kümnest ühedeistmest** tulõd siält ää (SAA)
(3) olliva **puust** tettu õkva egalüttel endäl (TAR)
(4) no miss meije **lapsepolvest** sis ikke tian rääkkida (RAN)

Keskpunkti lähedal paikneb ka lähtekoht sündmuselt või organisatsioonist (sundk), mida võib pidada kohatähenduse alaliigiks ja on selles töös teistest kohatähendustest eraldatud. Sageduserinevused eri murrete vahel võivad olla tingitud analüüsitud tekstide temaatikast, näiteks mõnes tekstis on rohkem sündmuste kirjeldusi, kuid tegemist on siiski elatiivi tavalise funktsiooniga (näited 5–6).

(5) siss esä tul **karast** kodo (SET)
(6) **soeast** tulin välja s enamb Kunda vaprikku tüöle minu ei vuettu (RAN)

Püsiühendi (pu), resultatiivi (tul) ja terviku (ter) funktsioone kasutatakse nii graafiku põhjal kui ka andmete järgi kõige rohkem idamurdes (näited 7–9), põhjuse (poh) funktsiooni esineb üsna palju kirdemurdes (näide 10), kuigi ida- ja kirdemurde kogu sõnade arvud on kõikide murrete arvestuses ühed väiksemad.

(7) kas sa mõessad lugeda küsib ka este minu **käess** (IDA)
(8) vanemast poeast **Voldemarist** koolittas õppettaea (IDA)
(9) mina olen **peregonnast** kõige viimane ja nüid olen **kõigist** järäle jäänd (IDA)
(10) **ehmattamisest** lüöb kramp (KIR)

Samas teistes murretes esineb kõiki neid funktsioone vähe või mitte üldse, seega on need ka pilve keskpunktist kaugemal. Siin on tõenäoliselt tegemist murretevahelise erinevusega elatiivi kasutuses, näiteks kuna kirdemurdes kasutatakse seda käänet rohkem põhjuse tähenduses, siis võib eeldada, et teised murded kasutavad sama tähenduse väljendamiseks mõnda muud käänet või konstruktsiooni.

Funktsioonide puhul on näha, et need muutuvad mööda horisontaalset mõõdet sagedamini kasutatavatest funktsioonidest nendeni, mis esinevad kõnes harvem. Murrete lõikes mööda mõõtmeid selgeid gruppe ei eristu. See on huvitav tulemus, sest siin ei eristu näiteks selgelt põhja- ja lõunaeesti murrete rühmad. Traditsioonilised murdejaotused võtavad arvesse ainult häälдust, vormi ja sõnavara ning morfosüntaktilist[6] tasandist murrete jaotamisel üldjuhul ei arvestata. Saadud tulemus illustreerib, et erineva lingvistilise tasandi põhjal võib murrete rühmitus anda oluliselt teistsuguseid tulemusi.

5.2. Standardiseeritud andmed

Analüüsis kasutatud andmestik on standardiseeritud kahe baasi põhjal. Korpuse keskmise sõnade arvu baasil standardiseeritud andmestik on esitatud lisas 2. 100 000 sõna baasil standardiseeritud andmestik on esitatud lisas 3.

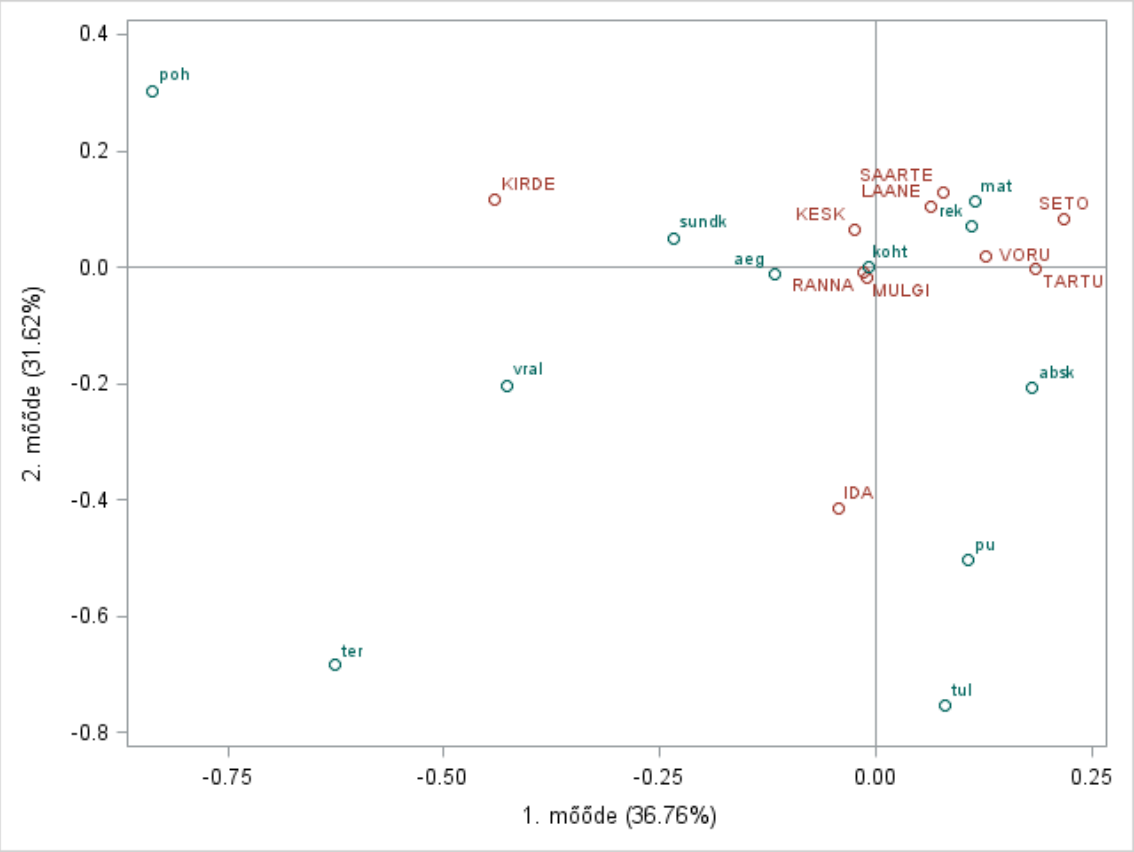
5.2.1. Standardiseerimisbaas korpuse keskmine sõnade arv

Korpuse keskmine sõnade arv on saadud liites kõikide murrete sõnade arvu ja jagades saadud tulemuse murrete arvuga. Selleks on, nagu eespool mainitud, 83 431 sõna. Keskmise sõnade arvuga standardiseeritud andmestikule rakendatud korrespondentsanalüüsi tulemused on esitatud joonisel 3. Kaks esimest mõõdet kirjeldavad ära üle 68% koguinertsist (vt ka lisa 4), seega standardiseerimine on vähendanud seda osa koguinertsist, mis jääb kahemõõtmelisel graafikul kirjeldamata.

Üldine pilt, võrreldes standardiseerimata andmetega, on suhteliselt sarnane: eristub üks suurem murrete grupp pilve keskpunktis ning ida- ja kirdemurre on taas eraldi. Samas on aga ka need kaks murret kesksele grupile lähenenud (ida horisontaalsel ja kirde

[6] Morfosüntaks on keeleteaduse haru, mis uurib grammatilisi kategooriaid ja keeleüksusi, millel on nii morfoloogilised (vormiõpetuslikud) kui ka süntaktilised (lauseõpetuslikud) omadused ning mida saab seetõttu defineerida nii morfoloogiliste kui ka süntaktilise kriteeriumide järgi.

vertikaalsel mõõtmel). See näitabki, et kui mõne nähtuse lihtsagedus on teistega võrreldes väga suur või väga väike, siis standardiseerimine aitab selliseid sageduserinevusi tasandada. Kõige sagedasemad funktsioonid – lähtekoht, algusaeg, materjal ja rektsioon – on samuti säilitanud oma positsiooni.



Joonis 3. Korrespondentsanalüüsi graafik korpuse keskmise sõnade arvu baasil standardiseeritud andmete põhjal.

Funktsioonidest eraldub teistest veelgi enam põhjus (poh), sest standardiseerimisjärgselt on selle sagedus kirdemurdes üle kahe korra suurem kui üheski teises murdes. Abstraktset lähtekohta (absk) esines lihtsagedustena palju kesk- ja saarte murdes, aga standardiseerimine suurendab selle osakaalu Tartu ja idamurdes, seega esineb see nende vahel ka graafikul (näited 11–12).

(11) aga ärä joo **meelest** lännu selle pikkä ajaga (TAR)
(12) maa len **raamattust** lugend (IDA)

Võrdlusalust (vral) esineb andmete järgi kõige rohkem ida-, kirde- ja Mulgi murdes (näited 13–15), samas väga vähe mujal, seega ongi see graafikul paigutunud nende kolme vahele. Terviku (ter) funktsiooni esineb standardiseeritud andmetes väga palju ida- ja kirdemurdes (näide 16), aga vähe mujal ja Tartu ning Seto murdes mitte ühtki korda, mida kirjeldab ka saadud graafik, kus see funktsioon on kirde- ja idamurdele kõige lähemal, samas siiski kaugel kõigist murretest.

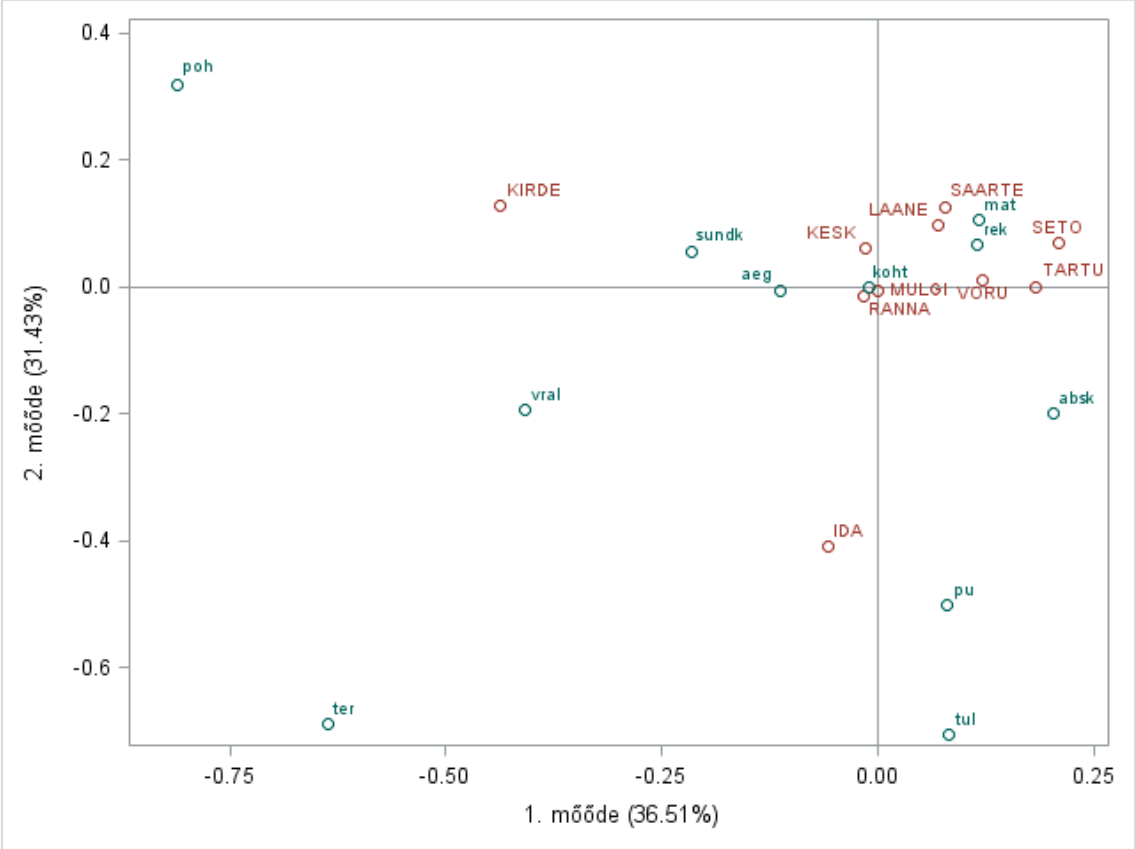
(13) oma **õdedest** oli uhkem ja ilusam (IDA)
(14) minu naabru, sie on **minust** vanemb inimine viel (KIR)
(15) mia pidi ikka alamb olema **temäst**, mia pidi ikki temä käsku tegemä (MUL)
(16) mina õlin viimane laps siis neist **neljast** lapsest (KIR)

5.2.2. Standardiseerimisbaas 100 000 sõna

Baasiga 100 000 standardiseeritud andmestiku põhjal tehtud korrespondentsanalüüsi graafik on kujutatud joonisel 4. Kuna 100 000 sõna on suurem kui korpuse keskmine sõnade arv, siis sellise baasiga standardiseerimine võimendab veelgi rohkem funktsioonide esinemissagedusi.

Saadud graafik on väga sarnane eelnevaga, sest on standardiseeritud enam-vähem sama baasiga. Seega ka enamus erinevusi standardiseerimata andmete ja 100 000 sõna baasil standardiseeritud andmete vahel kattuvad nendega, mis on juba mainitud eelnevas alapeatükis. Jooniselt on näha, et kaks mõõdet kirjeldavad ära 68% koguinertsist (vt ka lisa 5), mis on samuti peaaegu võrdne eelmises analüüsis saadud tulemusega.

Murded koonduvad veelgi enam vertikaalsel mõõtmel, kusjuures nii palju, et pilve keskpunktis asuvat murdegruppi ei ole enam kergesti võimalik omavahel eristada. See võib tähendada, et nendes murretes ongi elatiivi funktsioonide kasutamine väga ühesugune ning standardiseerimine võimaldab selle järelduseni jõudmist.



Joonis 4. Korrespondentsanalüüsi graafik 100 000 sõna baasil standardiseeritud andmete põhjal.

Lisaks on korrespondentsanalüüsi tulemustest näha, et standardiseerimise puhul koguinerts suureneb. Kui standardiseerimata andmete koguinerts on 0.06, siis korpuse keskmise sõnade arvu või 100 000 sõna baasil standardiseeritud andmetes on see vastavalt 0.0805 ja 0.0788 (vt lisa 1, 4–5). Koguinerts kirjeldab ridade ja veergude omavahelist korrelatsiooni: mida suurem on koguinerts, seda suurem on seos rea- ja veeruprofiilide vahel (Greenacre 1993: 30–31). Seega on standardiseeritud andmetes read ja veerud omavahel rohkem korreleeritud kui standardiseerimata andmetes. Samas on koguinertsid küllaltki väiksed, sest maksimaalseks koguinertsiks oleks mõõtmete arv (Greenacre 1993: 30), mis käesoleval juhul on kaks. Kokkuvõttes võib järeldada, et kuigi korrespondentsanalüüsi graafikute järgi võib murrete ja elatiivi funktsioonide vahel seoseid luua, ei pruugi need seosed olla väga tugevad.

Samuti tuleb arvestada, et saadud korrespondentsanalüüsi tulemused on suuresti mõjutatud töö autori funktsioonide valikust ning nende interpreteerimisest kodeerimisel, samas ka keelejuhtidest ning nende keelekasutusest. Kuigi need asjaolud ei mõjuta järgnevat analüüsi, kus on standardiseerimata ja standardiseeritud andmestikke omavahel võrreldud, siis on võimalik, et teistsugune kodeerimine oleks saavutanud ka erineva seose murrete ja elatiivi funktsioonide kasutuse vahel.

5.3. Tulemuste võrdlus

Korrespondentsanalüüside graafikute põhjal on näha küll väikest varieerumist standardiseerimata ja standardiseeritud andmetes, kuid üldine tulemus on erinevaid andmestikke kasutades sama. Järelikult ei saa korrespondentsanalüüsi graafikute põhjal veel väita, et standardiseerimine avaldaks analüüsi tulemustele mõju.

Andmestike täiendavaks analüüsimiseks on kasutatud hii-ruut-testi, millega saab hinnata, kas võrreldavad tunnused on omavahel sõltuvuses (Tooding 2015: 176). Pearsoni χ^2-statistik leitakse järgmise valemiga:

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{\left(n_{ij} - {n_{ij}}'\right)^2}{{n_{ij}}'},$$

kus n_{ij} on empiirilise jaotuse *i*-nda rea ja *j*-nda veeru sagedus, ${n_{ij}}'$ on teoreetilise jaotuse *i*-nda rea ja *j*-nda veeru sagedus ning k ja m on vastavalt ridade ja veergude arv andmestikus (Tooding 2015: 207). Teoreetilised sagedused arvutatakse valemiga

$${n_{ij}}' = n \cdot \frac{n_{i+}}{n} \cdot \frac{n_{+j}}{n},$$

kus n_{i+} on *i*-nda rea kogusumma, n_{+j} on *j*-nda veeru kogusumma ja n on andmestiku kogusumma (Tooding 2015: 207).

Hii-ruut-test kontrollib võrreldavate tunnuste sõltuvust:

$$H_0: tunnused\ on\ sõltumatud$$

$$H_1: tunnused\ on\ sõltuvad$$

Antud juhul väidab nullhüpotees, et murded ja elatiivi funktsioonide kasutus on üksteisest sõltumatud. Sisuka hüpoteesi kohaselt eksisteerib murrete ja funktsioonide kasutuse vahel seos.

Hii-ruut-testi kasutamiseks ei tohiks andmestikes esinevate sageduste väärtused olla liiga väiksed. Mitmete käsitluste (Parring jt 1997, Kilgarriff 2001, Tooding 2015 jt) järgi ei sobi seda testi tunnuste võrdlemiseks kasutada, kui lahtrite oodatavad sagedused on alla viie. Kasutatud andmestikes esineb küll mõningaid madalaid oodatavaid sagedusi, enamus neist on aga suuremad kui kolm, seega ei hakka need töö autori arvates testi tulemust oluliselt mõjutama.

χ^2-statistiku abil saab kontrollida tunnuste sõltuvuse hüpoteese, kuid sellega ei saa kirjeldada kahe tunnuse vahelise seose tugevust, sest statistiku muutumispiirkond sõltub tabeli mõõtmetest. Sõltuvuse tugevuse kirjeldamiseks on vaja normeeritud seosekordajat, mille väärtused asuksid teataval lõigul (Parring jt 1997: 217). Statistikapakett SAS väljastab erinevaid hii-ruut-statistiku standardiseerimisel saadud seosekordajaid, millega saab võrrelda erisuguse kujuga ja esinemissagedustega tabeleid (Tooding 2015: 207). Siin töös on andmestike võrdlemiseks kasutatud Craméri seosekordajat V, mis leitakse valemiga

$$V = \sqrt{\frac{\chi^2}{n(L-1)}},$$

kus $L = \min(k, m)$. Craméri kordaja mõõdab risttabeli lähedusastet sellisele tabelile, kus kehtib ühene vastavus rea- ja veerutunnuste väärtuste vahel ehk kus ühe tunnuse iga väärtusega esineb koos üks teise tunnuse väärtus (Tooding 2015: 208). Mida suurem on kordaja väärtus, seda tugevam on seos kahe tunnuse vahel. Kui sagedustabelis on ridade

ja veergude arv võrdne, asub kordaja V väärtus lõigus [0,1], vastasel juhul võib selle kordaja väärtus olla ühest suurem (Parring jt 1997: 218).

Seega kui hii-ruut-testi tulemused ja Craméri seosekordaja on nii standardiseerimata kui standardiseeritud andmete põhjal samad või sarnased, saaks väita, et andmete standardiseerimisel ei ole selle analüüsi tulemustele mõju olnud.

Olgu olulisuse nivoo $\alpha = 0.05$. Tarkvara SAS kasutades on standardiseerimata andmetel läbi viidud hii-ruut-test. Selle tulemusel on hii-ruut-statistiku väärtuseks saadud $\chi^2 = 353.82$ ja olulisustõenäosuseks $p < 0.0001$ (vt lisa 6). Seega tuleb vastu võtta sisukas hüpotees, et murded ja elatiivi funktsioonide kasutus on omavahel sõltuvad.

Korpuse keskmise sõnade arvu baasil standardiseeritud andmete korral on $\chi^2 = 553.66$ ja 100 000 sõna baasil standardiseeritud andmete puhul $\chi^2 = 549.65$. Olulisustõenäosus on mõlemal juhul $p < 0.0001$ (vt lisa 7–8), järelikult ka nende andmestike korral kehtib sisukas hüpotees, mille kohaselt on murded ja funktsioonid seotud.

Craméri seosekordaja on standardiseerimata andmete korral väärtusega $V = 0.082$, korpuse keskmise sõnade arvu ja 100 000 sõna baasil standardiseeritud andmete puhul vastavalt $V = 0.104$ ja $V = 0.094$ (vt lisa 6–8). Nendest kordajatest järeldub, et murrete ja funktsioonide kasutuse vahel esineb nõrk seos. Sageduste standardiseerimisel sõltuvus tunnuste vahel suureneb, mida näitas ka korrespondentsanalüüsi koguinertside võrdlemine eelmises alapeatükis, kuid see suureneb väga vähesel määral.

Võrreldes hii-ruut-testi tulemusi kolmel eri juhul, on näha, et nendest saadud järeldused on väga sarnased. Murrete ja elatiivi funktsioonid on küll omavahel sõltuvad, kuid kõikide analüüside kohaselt on nende vahel nõrk seos. Järelikult ei ole sageduste standardiseerimine selle analüüsi tulemusi mõjutanud.

Sellest ei saa aga lõplikult järeldada, et standardiseerimine on kõikides statistilistes analüüsides ebavajalik. On võimalik, et mõne muu meetodiga, mis ei ole

korrespondentsanalüüs, oleksid saadud tulemustes suuremad erinevused ilmnenud. Selleks, et väita, et standardiseerimine statistilise analüüsi tulemustele mõju ei avalda, tuleks teha täiendavaid analüüse.

Käesolevas alapeatükis läbi viidud analüüsi tulemused näitavad, et bakalaureusetööle püstitatud hüpotees tuleb selle konkreetse analüüsi põhjal ümber lükata. Seega peab järeldama, et sageduste standardiseerimine ei mõjuta oluliselt korrespondentsanalüüsi tulemusi.

KOKKUVÕTE

Käesoleva bakalaureusetöö eesmärk oli välja selgitada, kas sageduste standardiseerimisel on korrespondentsanalüüsi tulemustele mõju. Sageduste standardiseerimine on meetod, mida kasutatakse juhul, kui võrreldavad tekstid, korpused või ka inimgrupid on ebavõrdsete suurustega. Standardiseerimise puhul korrutatakse keelenähtuse suhteline sagedus baasiga, mis on valitud vastavalt tekstide mahule. Töö autori hüpotees oli, et sageduste standardiseerimine mõjutab korrespondentsanalüüsi tulemusi.

Töö esimeses peatükis anti ülevaade sageduste standardiseerimise meetodist. Teine peatükk kirjeldas elatiivi ajalugu ja selle funktsioone. Kolmandas peatükis tutvustati andmestikku, millel standardiseerimise mõju uuriti ja mis kirjeldas elatiivi funktsioonide kasutamist kümnes eesti murdes. Neljas peatükk selgitas kasutatud meetodit – korrespondentsanalüüsi.

Viiendas peatükis viidi läbi analüüs võrdlemaks standardiseeritud andmestikke standardiseerimata andmestikuga. Kasutatud andmeid standardiseeriti kahe eri baasiga: 83 431 sõna (korpuse keskmine sõnade arv) ja 100 000 sõna. Kõikidel andmestikel rakendati korrespondentsanalüüsi, mille tulemusel saadud graafikud erinesid üksteisest vaid vähesel määral. Seega ei olnud nende põhjal võimalik vastu võtta otsust, et saadud tulemused on üksteisest oluliselt erinevad.

Edasi võrreldi andmestikke hii-ruut-testi abil, millega kontrolliti, kas murrete ja elatiivi funktsioonide kasutuse vahel eksisteerib seos ning kui tugev see seos on. Kõigi kolme andmestiku korral selgus, et murrete ja funktsioonide vahel esineb nõrk sõltuvus. Craméri seosekordaja oli kõikide testide korral $V \approx 0.1$. Kuigi standardiseerimisel kordaja pisut suurenes, olid analüüside järeldused siiski samad. Seega selle konkreetse analüüsi tulemusi sageduste standardiseerimine ei mõjutanud.

Töö lisaeesmärgiks oli uurida elatiivi funktsioonide kasutuse varieerumist murrete lõikes. Korrespondentsanalüüsi graafikutelt ilmnes, et ida- ja kirdemurdes kasutatakse neid funktsioone teistest murretest erinevalt (idamurdes enam püsiühendi ja resultatiivi funktsioonis, kirdemurdes rohkem põhjuse funktsioonis). Ülejäänud murded paigutusid graafikul pilve keskpunkti lähedusse, seega neis murretes kasutatakse erinevaid elatiivi funktsioone pigem ühtlasema sagedusega. Samas oli ka korrespondentsanalüüsi koguinerts üsna väike (kõigi kolme analüüsi puhul väiksem kui 0.1) ehk murrete ja elatiivi funktsioonide kasutamissageduste vahel ei esinenud tugevaid seoseid.

Bakalaureusetöö eesmärk uurida, kas standardiseerimisel on korrespondentsanalüüsi tulemustele mõju, sai seega täidetud. Tööle püstitatud hüpotees tuli selle konkreetse analüüsi põhjal ümber lükata ning võeti vastu sisukas hüpotees, mille kohaselt standardiseerimine analüüsi tulemustele mõju ei avalda. Saadud tulemus kehtib aga ainult selles töös kasutatud andmestiku ja analüüsi korral. Selleks, et veenduda, et standardiseerimine tõepoolest statistiliselt oluliselt kvantitatiivse analüüsi tulemusi ei mõjuta, tuleks läbi viia täiendavaid analüüse.

LÜHENDID

Elatiivi funktsioonid

aeg – algusaeg

absk – abstraktne lähtekoht

e – viga kodeeringus

f – fraasi osa

koht – lähtekoht

mat – materjal

muu – muu funktsioon

põh/poh – põhjus

pü/pu – püsiühend

rek – rektsioon

sündk/sundk – lähtekoht sündmuselt või organisatsioonist

ter – tervik

tul – resultatiiv

vral – võrdlusalus

Muud lühendid

KES – keskmurre

KIR – kirdemurre

IDA – idamurre

LÄÄ – läänemurre

MUL – Mulgi murre

RAN – rannamurre

SAA – saarte murre

SET – Seto murre

TAR – Tartu murre

VÕR – Võru murre

KIRJANDUS

Adolphs, Svenja 2006. Introducing Electronic Text Analysis. A practical guide for language and literary studies. London/New York: Routledge.

Ahrens, Eduard 2003 [1953]. Eesti keele Tallinna murde grammatika. Teine osa: lauseõpetus. Tõlk. Kristi Mets, Kristiina Rebane, Mailis Salvet, toim. Kristiina Ross. Tallinn: Eesti Keele Sihtasutus.

Biber, Douglas 1988. Variation across speech and writing. Cambridge: Cambridge University Press.

Biber jt = Biber, Douglas, Susan Conrad, Randi Reppen 1998. Corpus linguistics. Investigating language structure and use. Cambridge: Cambridge University Press.

EKG = Erelt, Mati, Reet Kasik, Helle Metslang, Henno Rajandi, Kristiina Ross, Henn Saari, Kaja Tael, Silvia Vare 1995. Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

EKK = Erelt, Mati, Tiiu Erelt, Kristiina Ross 2007. Eesti keele käsiraamat. Kolmas, täiendatud trükk. Tallinn: Eesti Keele Sihtasutus.

EKI 2014 = Eesti Keele Instituudi kohanimeandmebaasi kihelkonnapiiride andmestik.

EMK 2015. Eesti murrete korpus; http://www.keel.ut.ee/et/keelekogud/murdekorpus. Vaadatud 13.03.2016.

Greenacre, Michael 1984. Theory and Applications of Correspondence Analysis. London: Academic Press.

Greenacre, Michael 1993. Correspondence Analysis in Practice. London: Academic Press.

Kilgarriff, Adam 2001. Comparing corpora. – Corpus Linguistics. Critical Concepts in Linguistics. Volume II. Toim. Wolfgang Teubert, Ramesh Krishnamurthy. London/New York: Routledge, 232–263.

Mäearu, Sirje 2011. Valik rektsioone. Tartu: Keelehooldekeskus.

Pajusalu jt = Pajusalu, Karl, Tiit Hennoste, Ellen Niit, Peeter Päll ja Jüri Viikberg 2009. Eesti murded ja kohanimed. 2., täiendatud trükk. Toim. Tiit Hennoste. Tallinn: Eesti Keele Sihtasutus.

Parring jt = Parring, Anne-Mai, Mare Vähi, Ene Käärik 1997. Statistilise andmetöötluse algõpetus. Tartu: Tartu Ülikooli Kirjastus.

Pärna, Kalev 1993. Correspondence Analysis: An Introduction and Some Examples. Stockholm: Stockholm University, Department of Statistics.

Ross, Kristiina 1997. Kohakäänded Georg Mülleri ja Heinrich Stahli eesti keeles. – Pühendusteos Huno Rätsepale. Tartu Ülikooli eesti keele õppetooli toimetised 7. Tartu: Tartu Ülikooli Kirjastus, 184–201.

Rätsep, Huno 1979. Eesti keele ajalooline morfoloogia II. Tartu: Tartu Riiklik Ülikool, eesti keele kateeder.

Tooding, Liina-Mai 2015. Andmete analüüs ja tõlgendamine sotsiaalteadustes. Teine, täiendatud väljaanne. Tartu: Tartu Ülikooli Kirjastus.

Uiboaed, Kristel 2013. Verbiühendid eesti murretes. Tartu: Tartu Ülikooli Kirjastus.

Wiedemann, Ferdinand Johann 2011 [1875]. Eesti keele grammatika. Tõlk. Heli Laanekask, toim: Ellen Niit. Tallinn: Eesti teaduste Akadeemia Emakeele Selts.

LISAD

Lisa 1. SASi väljavõte inertside tabelist standardiseerimata andmete korral

Inertia and Chi-Square Decomposition					
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	7 14 21 28 35 ----+----+----+----+----+---
0.14111	0.01991	117.393	33.18	33.18	**********
0.13027	0.01697	100.064	28.28	61.46	**********
0.09124	0.00833	49.088	13.87	75.33	**********
0.08520	0.00726	42.799	12.10	87.43	*********
0.06454	0.00417	24.559	6.94	94.37	*****
0.04755	0.00226	13.328	3.77	98.14	***
0.02577	0.00066	3.916	1.11	99.24	*
0.02109	0.00044	2.623	0.74	99.98	*
0.00302	0.00001	0.054	0.02	100.00	
Total	0.06001	353.824	100.00		
Degrees of Freedom = 90					

Lisa 2. Korpuse keskmise sõnade arvu baasil standardiseeritud andmestik

	ida	kesk	kirde	Mulgi	ranna	saarte	Tartu	lääne	Seto	Võru	kokku
absk	17	19	4	11	8	10	17	4	2	15	107
aeg	37	37	35	29	27	30	14	34	19	15	277
koht	317	258	289	284	452	320	210	278	213	203	2824
mat	116	137	109	99	123	171	102	171	132	99	1259
põh	7	20	47	12	13	10	1	12	2	7	131
pü	61	18	12	17	21	15	23	17	23	11	218
rek	59	73	60	81	73	66	66	72	68	72	691
sündk	9	8	16	8	11	10	10	5	6	2	85
ter	18	5	11	3	5	2	0	3	0	2	49
tul	17	3	0	8	8	1	0	5	2	7	51
vral	22	8	25	14	11	8	8	10	2	7	115
kokku	680	586	608	566	752	643	451	611	469	441	5807

Lisa 3. 100 000 sõna baasil standardiseeritud andmestik

	ida	kesk	kirde	Mulgi	ranna	saarte	Tartu	lääne	Seto	Võru	kokku
absk	20	23	4	31	10	12	20	5	3	19	1297
aeg	44	44	42	35	33	37	17	40	23	19	334
koht	380	309	346	340	542	384	252	333	255	243	3384
mat	139	165	130	118	147	1061	122	205	158	119	1509
põh	9	24	57	14	15	13	2	14	3	9	160
pü	73	22	15	20	25	18	27	21	28	13	262
rek	71	88	71	98	87	80	79	87	82	87	830
sündk	11	9	19	9	14	12	12	6	8	3	103
ter	22	6	13	3	6	2	0	4	0	3	59
tul	20	4	0	9	10	2	0	6	3	9	63
vral	27	10	29	17	14	10	9	12	3	9	140
kokku	816	704	726	676	903	776	540	733	566	533	6973

Lisa 4. SASi väljavõte inertside tabelist keskmise korpuse sõnade arvu baasil standardiseeritud andmete korral

Inertia and Chi-Square Decomposition					
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	7 14 21 28 35 ----+----+----+----+----+---
0.17198	0.02958	171.764	36.76	36.76	**************************
0.15951	0.02544	147.756	31.62	68.38	**********************
0.10011	0.01002	58.203	12.46	80.83	*********
0.08738	0.00764	44.337	9.49	90.32	*******
0.06443	0.00415	24.107	5.16	95.48	****
0.04808	0.00231	13.425	2.87	98.35	**
0.02678	0.00072	4.164	0.89	99.24	*
0.02428	0.00059	3.425	0.73	99.97	*
0.00450	0.00002	0.118	0.03	100.00	
Total	0.08047	467.299	100.00		
Degrees of Freedom = 90					

Lisa 5. SASi väljavõte inertside tabelist 100 000 sõna baasil standardiseeritud andmete korral

Inertia and Chi-Square Decomposition					
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	7 14 21 28 35 ----+----+----+----+----+---
0.16966	0.02878	200.703	36.51	36.51	**************************
0.15739	0.02477	172.733	31.43	67.94	**********************
0.10093	0.01019	71.034	12.92	80.86	*********
0.08725	0.00761	53.085	9.66	90.52	*******
0.06287	0.00395	27.559	5.01	95.54	****
0.04788	0.00229	15.985	2.91	98.44	**
0.02562	0.00066	4.576	0.83	99.28	*
0.02344	0.00055	3.831	0.70	99.97	
0.00453	0.00002	0.143	0.03	100.00	
Total	0.07883	549.649	100.00		
Degrees of Freedom = 90					

Lisa 6. SASi väljavõte hii-ruut-testi tulemustest standardiseerimata andmete korral

Statistic	DF	Value	Prob
Chi-Square	90	353.8238	<.0001
Likelihood Ratio Chi-Square	90	321.3921	<.0001
Mantel-Haenszel Chi-Square	1	3.0702	0.0797
Phi Coefficient		0.2450	
Contingency Coefficient		0.2379	
Cramer's V		0.0817	

Lisa 7. SASi väljavõte hii-ruut-testi tulemustest keskmise korpuse sõnade arvu baasil standardiseeritud andmete korral

Statistic	DF	Value	Prob
Chi-Square	90	553.6601	<.0001
Likelihood Ratio Chi-Square	90	502.9356	<.0001
Mantel-Haenszel Chi-Square	1	0.0883	0.7663
Phi Coefficient		0.3104	
Contingency Coefficient		0.2964	
Cramer's V		0.1035	

Lisa 8. SASi väljavõte hii-ruut-testi tulemustest 100 000 sõna baasil standardiseeritud andmete korral

Statistic	DF	Value	Prob
Chi-Square	90	549.6489	<.0001
Likelihood Ratio Chi-Square	90	507.0108	<.0001
Mantel-Haenszel Chi-Square	1	6.6111	0.0101
Phi Coefficient		0.2808	
Contingency Coefficient		0.2703	
Cramer's V		0.0936	

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Hanna Pook,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Sageduste standardiseerimise mõju korrespondentsanalüüsi tulemustele eesti murretes esinevate elatiivi funktsioonide näitel“, mille juhendajad on Mare Vähi ja Kristel Uiboaed;

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace'is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;

1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace'i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 29.04.2016