Tartu Ülikool

Humanitaarteaduste ja kunstide valdkond

Eesti ja üldkeeleteaduse instituut

Hanna Pook, Helen Kaljumäe

RELATIIVPRONOOMENI *KES* KASUTUS EESTI MURRETES

Projekt aines „Korpuslingvistika“

Tartu 2017

Sisukord

Sissejuhatus

Selles projektis on uurimise all relatiivse asesõna *kes* kasutus kümnes eesti murdes. See teema on valitud seotult ühe autori magistritööga, milles uuritakse relatiivpronoomenite *mis* ja *kes* kasutust elusatele ja elututele entiteetidele viitamiseks. Kuigi kirjakeeles on üldiselt *mis* seotud elutute entiteetidega ja *kes* seotud elusate entiteetidega, siis mõnes murdes võib elututele entiteetidele viidata ka asesõnaga *kes*. Lisaks pronoomeni *kes* kasutuse varieerumisele murretes on proovitud leida ka esialgseid vastuseid magistritöö uurimisküsimustele.

Seega on uurimuse eesmärgiks saada teada, kui palju esineb murdetekstides pronoomenit *kes*, milline on selle kasutusviis ning kas ja kuidas see erineb vaadeldavate murrete vahel. Uurimuse kaudu on võimalik leida, millistes vormides ning konstruktsioonides murretes üldse sõna *kes* leidub ning kuivõrd need kasutusviisid on piirkonniti erinevad.

1. Andmestik ja töökäik

Töö uurimiseesmärgi täitmiseks vajalik materjal on kogutud eesti murrete korpuse morfoloogiliselt märgendatud tekstide osast. See tähendab, et nendes litereeritud suulise kõne tekstides on igale tekstisõnale lisatud märksõna, sõnaliik, morfoloogiline info, vajadusel ka tähendus ja fraas. Morfoloogiliselt märgendatud tekstid on XML-formaadis.

Töös on uurimise all on kõik eesti murded: ida-, kesk-, lääne-, saarte, kirde-, ranna-, Mulgi, Tartu, Setu ja Võru murre. Iga keelejuhi tekst on eraldi XML-failis, seega koosneb iga murde materjal paljudest XML-failidest. Kasutatud materjal on seisuga 8.11.2016. Selles töös on iga murde materjali uuritav osa koondatud üheks failiks, seega keelejuhte eraldi vaadeldud ei ole, kuid saadud failid on jäetud murdeti eraldi, et murdeid oleks võimalik omavahel võrrelda.

Siinse uurimuse jaoks on esialgu koostatud iga murde kohta kaks andmestikku. Esimesed neist on moodustatud skriptiga `kes_loend.sh`, mille abil on murdetekstidest välja võetud kõik pronoomeni *kes* esinemisjuhud koos nende morfoloogilise infoga. Edasi on skriptiga `kes_sagedusloend.sh` iga murde kohta koostatud sagedusloend pronoomeni *kes* käänetest. Nende sagedusloendite tulemusi on kirjeldatud peatükis 2.1.

Teised andmestikud on loodud skriptiga `kes_uhendid.sh`, mis koostab loendid tekstis üksteisele järgnevatest sõnapaaridest, mille üheks liikmeks on pronoomeni *kes* ükskõik milline vorm. Selline andmestik sisaldab nii sõnapaare, kus *kes* on esimene liige, kui ka sõnapaare, kus *kes* on teine liige. Kuna relatiivpronoomen *kes* laiendab üldiselt endale eelnevat sõna, siis on eeldatud, et see seos kehtib ka uuritavate murdetekstide puhul. Seega on skriptiga `kes_eelnev_sona.sh` moodustatud uus sõnapaaride loend, kuhu on alles jäetud need eelmise loendi sõnapaarid, mille teine liige pronoomen *kes*. Siinkohal tuleb mainida, et suulises kõnes ei pruugi pronoomenile eelnev sõna alati olla just see sõna, millele *kes* viitab, seega võib uurimuse tulemustes esineda mõningaid vigu. Selliste vigade vähendamiseks on andmestikku alles jäetud vaid need sõnapaarid, kus pronoomenile eelnev sõna on nimisõna või asesõna, millele pronoomen *kes* ka päriselt viidata saab. Samuti saab vaid nimi- ja asesõnade puhul määrata sõna elusust, mis ongi selle uurimuse üheks eesmärgiks. Nende andmestike tulemused on kirjas peatükis 2.2.

Lõpuks on vastavalt töös leitud tulemustele arvutatud ühes murdes kahe sõnapaari jaoks ka t-skoor ja z-skoor, et määrata, milline on nende sõnapaaride sõnadevahelise seose tugevus ja kas see ka projektis eelnevalt saadud tulemusi kinnitab. Sõnadevahelise seose tugevuse mõõdikute analüüs on peatükis 2.3.

2. Analüüs ja tulemused

Skriptidega loodud kolme erineva andmestiku põhjal uuritakse järgnevalt eesti murretes relatiivpronoomeni *kes* käänete esinemist ja sõnu, millele pronoomen *kes* viitab.

2.1. Pronoomeni kes käänded

Pronoomeni *kes* kasutamist murretes võrreldi sagedusloendite abil, mis koostati tekstimaterjalides esinenud *kes*-i käänetest igale murdele. Andmete saamiseks ja sagedusloendi koostamiseks kasutati skripte `kes_loend.sh` ja `kes_sagedusloend.sh`.

Murrete sagedusloendeid uurides selgub, et sagedasim kääne on ainsuse nominatiiv (esineb kümnes murdes), teisel kohal esineb enim ainsuse adessiiv (üheksas murdes, v.a idamurdes), järgnevad ainsuse komitatiiv (viies murdes – Mulgi, Tartu, ida-, kesk- ja läänemurdes), ainsuse genitiiv (viies murdes – Mulgi, Setu, ida-, lääne- ja rannamurdes), ainsuse allatiiv (kolmes murdes – Setu, Võru ja kirdemurdes).

Sagedusloendeid iseloomustab see, et kõige sagedasem kääne (ehk ainsuse nominatiiv) on suure sagedusega (nt kirdemurdes esinenud 106 korda), kuid sealt edasi väheneb sagedus isegi mitmekordselt ning vahed eri käänete vahel on edasi paljudes murretes juba üsna väiksed (nt kirdemurdes on sageduselt teist ehk ainsuse adessiivi esinenud 13 korda, sellele järgneb üheksa korda esinev ainsuse genitiiv). Suurem vahe sageduselt teise ja kolmanda käände vahel on vaid Võru, Setu, saarte ja läänemurdes, mis on tingitud ainsuse adessiivi suuremast sagedusest nendes murretes.

Koostatud sagedusloenditest on näha ka seda, et kõige vähem erinevaid käändeid on kasutatud Mulgi murdes (kuus käänet), kõige rohkem aga saarte ja rannamurdes (kümme käänet). Tuleb aga siiski arvestada, et sagedusloendite lõpus leidub käändeid, mida on kasutatud vaid kord või paar ning seetõttu seda erinevust niivõrd tähenduslikuks lugeda ei tasu.

Pronoomeni sagedusloenditesse on jäetud alles nii ainsuse kui ka mitmuse vormid, et näha, kas mõnes murdes kasutatakse pronoomenit *kes* ka mitmuses, kuna kirjakeeles esinevad *kes*-i mitmuse vormid väga harva. Materjali eelnevalt uurides on näha, et selles on mitmuse märgend lisatud ka nendele kes vormidele, mis ise on ainsuslikud, kuid mis viitavad mitmuslikule sõnale (nt läänemurdes *teised, kis,* keskmurdes *esimihed, kiss*, Võru murdes *noorembaq, kess*). Seega kuigi sagedusloenditest saab näiteks välja lugeda, et mitmuse nominatiivi leidub kõikides murretes peale Mulgi murde, on tegelikult mitmusliku nominatiivi vorm ainult saarte murdes kujul *kessid*. Tartu murdes esinev mitmuse adessiiv, saarte murdes esinev komitatiiv ja rannamurdes esinevad mitmuse komitatiiv ja genitiiv on samuti tekstides ainsuslikud. *kes*-i mitmust leidubki veel ainult saarte murdes, kus mitmuse genitiiv on kujul *killette*.

2.2. Sõnapaarid pronoomeniga kes

Selliste sõnade leidmiseks, millele *kes* eeldatavalt viitab, koostati loendid *kes*-ile eelnevatest nimi- ja asesõnadest. Sõnapaaride leidmiseks andmestikust ja neist sagedusloendite koostamiseks kasutati skripte `kes_uhendid.sh` ja `kes_eelnev_sona.sh`.

Sagedusloendeid võrreldes selgub, et kõige sagedamini viitab *kes* asesõnale *see* ja selle mitmuse vormile *need*. Seejuures on näha ka piirkonnast tulenevat erinevust: Tartu, Võru ja Setu murdes on *see* ja *need* asemel kasutatud pronoomeneid *too* ning *nood*. Üldiselt viitavad asesõnad *ta* ja *nad* elusatele entiteetidele ning asesõnad *see* ja *need* elututele entiteetidele, kuid kuna ka kirjakeeles kasutatakse tihti neid asesõnu vastupidiselt, siis ei ole selles töös asesõnu *see* ja *need* (või lõunaeesti murrete *too* ja *nood*) täiesti elututeks loetud.

Suur osakaal on loendites üks kord esinevatel sõnadel. Rohkem kui kord esinevaid sõnu on näiteks saarte murdes 19, aga Setu murdes vaid neli. Analüüsis on täpsema uurimise alt jäetud kõrvale vaid korra pronoomeni *kes* ees kasutatud sõnad, sest nende puhul võis *kes*-i kasutus olla juhuslik. Mõned üks kord esinenud sõnad ei pruugi olla ka need, millele *kes* viitab, kuna suulises kõnes ei ole relatiivpronoomen alati täpselt viidatava sõna järel. Vaadeldes rohkem kui korra esinenud sõnu, võib öelda, et *kes* viitab murretes mitmetele asesõnadele, mis

asendavad elusat entiteeti (nt *mõni, keegi, teine, kõik*), ning nimisõnadele, mis tähistavad elusolendit (nt *inimene, mees, sepp, laps, loom, naine*).

Mitmes murdes leidub aga ka selliseid sõnu, mis ei tähista elusaid entiteete. Rohkem kui korra murretes esinevad elutud sõnad on loetletud tabelis 1. Sealjuures on siia tabelisse kirja pandud ainult need sõnad, mida kasutatakse prototüüpselt elutuna (kuigi mõned neist võivad teatud kontekstis väljendada ka elusaid entiteete).

Tabel 1. Rohkem kui ühe korra esinenud sõnad murrete kaupa.

Murre	Elutud sõnad
Idamurre	*käsi, raha*
Keskmurre	*lõng, kool*
Läänemurre	*asi, tükk*
Saarte murre	*miski, võrk, osa, põrand, küla, kala*
Rannamurre	*rubla, paat*
Kirdemurre	-
Mulgi murre	-
Tartu murre	-
Võru murre	*tare, miski, asi*
Setu murre	-

Need tulemused näitavad, et *kes* võib teatud murretes viidata ka elututele entiteetitele. Saadud andmete kohaselt kasutatakse pronoomenit *kes* sellistes konstruktsioonides kõige rohkem põhjaeesti murderühmas, seega ida-, kesk-, lääne- ja saarte murdes. Sisendtekste lähemalt uurides selgub, et tabelis toodud elututele sõnadele on pronoomen *kes* viidanud tegelikult ainult kesk- ja läänemurdes (näited 1–3). Teiste sõnade puhul on tegemist lihtsalt suulises kõnes pronoomeni ette paigutunud sõnadega.

1. vat see pidi nii pieenikke ja (.) ilus olema see (.) **lõng kellest** kootti (.) kellest ned linad tehti sakstelle (KESKMURRE)

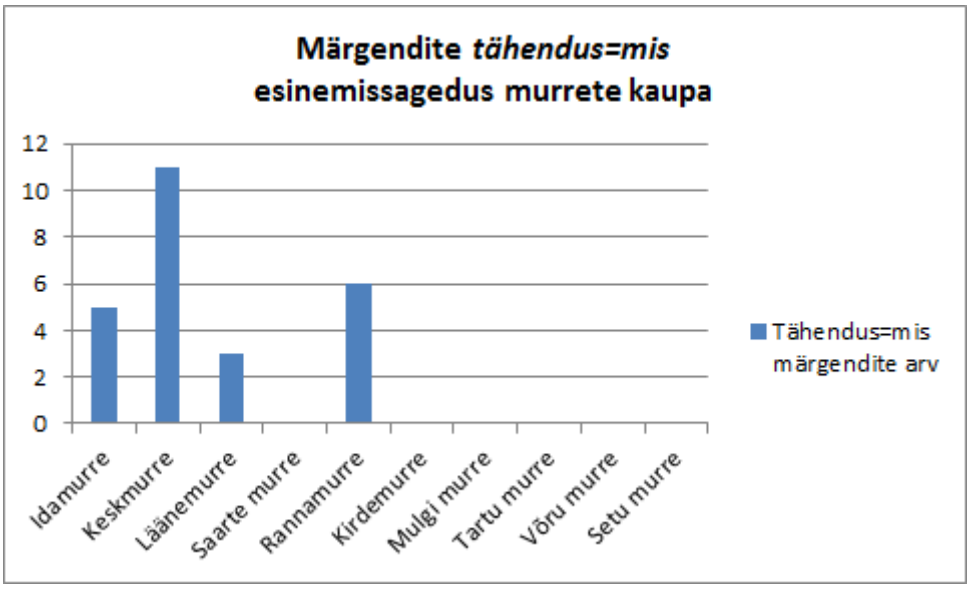
2. sial kedratti (...) **lõnga** (...) **kellest** riijet tehti (KESKMURRE)
3. nii tugevasti sai rautkividega ää täedettud (.) et=ta vauda eij=anna ja need onn kõik **munagivid=ja** (.) **väikset=tükkid** (.) **kellega**=see semendi kört on ühte valattud (LÄÄNEMURRE)

Sama olukord kehtib ka ranna- ja Võru murde puhul (näited 4-5), kus pronoomenile eelnev sõna pole see, millele pronoomen viitab. Varasemate teadmiste põhjal ei tohiks lõunaeesti murderühmas sellist konstruktsiooni üldse esineda.

4. neil miestel oli ikka suur **paat** (...) **kiss** seda nuotta vedasid neil oli ikka suur paat (RANNAMURRE)
5. muido niisamattõ **tarõhh kell**=oll kats **tarrõ kell** olõ=õss kattõ tarrõ (VÕRU MURRE)

Sellest saab järeldada, et antud uurimuseks püstitatud ülesanne ei ole selline, mida saaks kergesti skriptide abil lahendada. Andmed tuleks siiski inimesel mingil määral käsitsi kontrollida ja analüüsida.

Sisendfaile uurides on näha, et mõnes neist on sellised olukorrad, kus relatiivpronoomenit *kes* kasutatakse elutule entiteedile viitamiseks, märgendatud veel lisaks morfoloogilises infos märgendiga *tähendus=mis*. Tööd alustades otsustati need märgendid kõrvale jätta, sest neid on lisatud väga ebaühtlaselt, ilmselt ainult üksikute litereerijate poolt. Selleks aga, et kontrollida, kas eelnevas osas saadud tulemused paika peavad (st et põhjaeesti murderühmas kasutatakse relatiivpronoomenit *kes* elututele entiteetidele viitamiseks, aga lõunaeesti murderühmas sellist *kes*-i kasutust ei esine), on siiski leitud ka nende lisamärgendite sagedus murrete kaupa. Need lisamärgendid on eraldi igas murdes välja otsitud skriptiga `tahendus_mis.sh` ja nende sagedus on arvutatud skriptiga `tahendus_mis_arv.sh`. Need sagedused on esitatud joonisel 1.



Joonis 1. Märgendite *tähendus=mis* esinemissagedus murrete kaupa.

Kui eeldada, et *tähendus=mis* märgendi lisamine on olnud järjekindel, siis saab selle tabeli põhjal väita, et põhjaeesti murderühmas (eelkõige keskmurdes) kasutatakse tõesti relatiivpronoomenit *kes* relatiivpronoomeni *mis* asemel elututele entiteetidele viitamiseks. Samuti saab selle tabeli järgi kinnitust eeldus, et lõunaeesti murderühmas sellist konstruktsiooni ei esine. Kirderanniku murderühma kohta ühest järeldust teha ei saa, sest on võimalik, et *kes*-i kasutatakse elututele entiteetidele viitamiseks ainult rannamurdes, aga mitte kirdemurdes, kuid on ka võimalik, et kirdemurde tekstidele ei ole lihtsalt vastavat märgendit lisatud. Kuna aga on teada, et see märgendus on tekstides olnud ebaühtlane, siis usaldusväärsemate tulemuste jaoks peaks tegema täiendavaid analüüse.

2.3. z-skoor ja t-skoor

Kuna eelnevate tulemuste järgi tundub, et kõige rohkem kasutatakse *kes*-i elututes konstruktsioonides keskmurdes, siis on edasi otsustatud uurida täpsemalt just seda murret. Selleks on leitud sõnadevahelise seose tugevuse mõõdikud ehk z- ja t-skoorid sõnapaaride „see+kes” ja „see+mis” jaoks. Olgugi, et sõna *see* kasutatakse ka kirjakeeles tihti elusatele entiteetidele viitamiseks, on selle prototüüpsem eesmärk siiski viidata elututele entiteetidele. Seetõttu ongi uuritud, kas keskmurdes, kus relatiivpronoomenit *kes* kasutatakse tihti elututes konstruktsioonides, on sõnadevaheline seos sõnapaari „see+kes” puhul tugevam kui sõnapaari „see+mis” puhul. Sõna *see* on valitud selle pärast, et see on alapeatükis 2.2. leitud keskmurde

loendites kõige sagedasem sõna ja seega on tõenäolisem, et selle sõnaga saab paikapidavamaid tulemusi kui mõne muu väga selgelt elutu, aga harva esineva sõnaga.

Sõnapaari z-skoori ja t-skoori leidmiseks on vaja järgmiseid suuruseid: sõnapaari koosesinemise sagedus valimis (O), ühe sõnapaari kuuluva sõna sagedus ($f1$), teise sõnapaari kuuluva sõna sagedus ($f2$), valimimaht (N) ja sõnapaari koosesinemise teoreetiline sagedus (E), mis arvutatakse valemiga

$$E = \frac{f1 \cdot f2}{N}.$$

Lõpuks leitakse sõnapaari z-skoor valemiga

$$z = \frac{O - E}{\sqrt{E}}$$

ja sõnapaari t-skoor valemiga

$$t = \frac{O - E}{\sqrt{O}}.$$

Arvutatavad skoorid näitavad, kui tugev on sõnadevaheline seos: mida suurem on saadud seose tugevuse skoor, seda tugevamaks võib seost pidada. z- ja t-skoor võib olla nii positiivne kui ka negatiivne. Positiivne skoor tähendaks, et sõnapaari sõnad on omavahel seotud, negatiivse skoori korral seos puuduks. t-skoor on täpsem olukorras, kus sõnapaari koosesinemise sagedus on väike (mõnede autorite arvates peaks z-skoori kasutamiseks olema $E > 9$, kuid enamik neist nõustub, et t-skoor on vajalik olukorras, kus $E < 1$) (Evert 2007: 1225–1227).

Kõigepealt on arvutatud sõnapaari „see+kes" z- ja t-skoori jaoks vajalikud suurused. Keskmurde sõnade arv on leitud skriptiga `korpusesuurus.sh` ja selleks on $N = 146\,363$ sõna. Sellest arvust on välja jäetud kõik vahemärgid (st pausid, arusaamatud sõnad jms) ja muu metainfo (st intervjueerija kommentaarid, poolikud sõnad jms). Sõna *see* esinemise arv keskmurdes on $f1 = 4098$ sõna ja sõna *kes* esinemise arv on $f2 = 386$ sõna. Need tulemused on leitud vastavalt skriptidega `see_arv.sh` ja `kes_arv.sh`. Sõnapaari koosesinemise sagedus on arvutatud skriptiga `kes_see_arv.sh` ja selleks on $O = 40$ sõna. Sealjuures on koosesinemise aknaks ainult üksteisele järgnevad sõnad. Nii väike aken on valitud seetõttu, et suulises kõnes on väga keeruline lause pikkust määrata, seega kui valida

aknaks ainult üksteisele järgnevad sõnad, vähendab see võimalust, et saadud sõnapaar on pärit erinevatest lausetest. Sõnapaari sõnade järjekord võib olla nii „see kes" kui ka „kes see". Sõnapaari teoreetiline koosesinemise sagedus on

$$E = \frac{386 \cdot 4098}{146\,363} = 10{,}81.$$

Seega sõnapaari „see+kes" z-skoor on

$$z = \frac{40 - 10{,}81}{\sqrt{10{,}81}} = 8{,}88$$

ja t-skoor on

$$t = \frac{40 - 10{,}81}{\sqrt{40}} = 4{,}62.$$

z- ja t-skoori arvutamiseks sõnapaarile „see+mis" tuleb sellised paarid kõigepealt andmestikust leida, sest seni on töös tegeletud vaid pronoomeni *kes* käänete ja sõnapaaridega. Seega skriptiga `mis_uhendid.sh` koostatakse esmalt tekstis üksteisele järgnevate sõnapaaride loendid, kus sõnapaari üheks osaks on pronoomeni *mis* ükskõik milline vorm.

Et leida z- ja t-skoori sõnapaarile „see+mis", on vaja arvutada sõna *mis* esinemise arv, milleks on $f3 = 915$ sõna (leitud skriptiga `mis_arv.sh`), ja sõnapaari koosesinemise sagedus, milleks on $O = 92$ sõna (leitud skriptiga `mis_see_arv.sh`). Ka selle sõnapaari puhul on aknaks ainult üksteisele järgnevad sõnad ja sõnapaari sõnade järjekord võib olla nii „see mis" kui ka „mis see". Sõnapaari teoreetiline koosesinemise sagedus on

$$E = \frac{915 \cdot 4098}{146\,363} = 25{,}62.$$

Sõnapaari „see+mis" z-skoor on

$$z = \frac{92 - 25{,}62}{\sqrt{25{,}62}} = 13{,}11$$

ja t-skoor on

$$t = \frac{92 - 25{,}62}{\sqrt{92}} = 6{,}92.$$

Kuna antud juhul on leitud teoreetilised sagedused piisavalt suured, on mõistlik võrrelda omavahel leitud z-skoore. Nende puhul võib väita, et kui $|z| > 3{,}29$, siis on tegemist olulise

seosega (Evert 2007: 1227). Saadud tulemused näitavad, et kuigi nii sõnapaari “see+kes” kui ka “see+mis” sõnade vahel on oluline positiivne seos, siis sõnapaari „see+kes” z-skoor on siiski madalam kui sõnapaari „see+mis” z-skoor (vastavalt 8,88 ja 13,11). Selle põhjal võib öelda, et keskmurdes on tugevam seos ikkagi sõnade *mis* ja *see* vahel. Seega olgugi, et keskmurdes kasutatakse relatiivpronoomenit *kes* elututele entiteetidele viitamiseks, ei ole see kasutus olulisem kui *mis*-iga elututele entiteetidele viitamine. Samas võib saadud tulemus olla seotud sellega, et elutuks sõnaks oli uurimise all asesõna *see*, mida on võimalik kasutada ka elusate entiteetide kohta.

Kokkuvõte

Töös uuriti relatiivpronoomeni *kes* kasutust kümnes eesti murdes. Vaadati, millistes vormides esineb murretes pronoomen *kes* ning missugustele sõnadele *kes* viitab. Tekstimaterjal koguti eesti murrete korpusest, töös kasutatud vajalike andmete saamiseks koostasid autorid skriptid.

Pronoomeni *kes* käänete võrdlemisel murdeti selgus, et sagedasimad käänded on ainsuse nominatiiv, mis esines kõigis kümnes murdes, ja ainsuse adessiiv, mida leidus üheksas murdes. Iseloomulik on seegi, et kui kõige enam esinev kääne on suure sagedusega, siis järgnevate käänete nii sagedused kui ka vahed on väiksemad. Lisaks käändele vaadati ka seda, kas *kes* võib mõnes murdes esineda ka mitmuses. Kuigi skriptidega saadud andmestiku põhjal jäi mulje, et neid vorme on kasutatud mitmeid kordi, siis tegelikkuses esines mitmuslikku pronoomenit *kes* vaid saarte murdes.

Selliste sõnade uurimiseks, millele kes eeldatavalt viitab, koostati loendid *kes*-ile eelnevatest nimi- ja asesõnadest. Sagedusloenditest võis näha, et *kes* viitab murretes mitmetele asesõnadele, mis asendavad elusat entiteeti, ning nimisõnadele, mis tähistavad elusolendit.

Mitmes murdes oli *kes*-i ees ka sõnu, mis ei tähista elusaid entiteete. Kõige enam leidus selliseid sõnu põhjaeesti murderühmas. Andmestikku kontrollides aga ilmnes, et tegelikult on elututele sõnadele *kes*-iga viidatud vaid kesk- ja läänemurdes ning teiste puhul oli tegu vaid pronoomeni ette paigutunud sõnadega. Algses andmestikus olevaid lisamärgendeid arvestades selgus siiski, et põhjaeesti murderühmas võib elutule viitamist esineda, lõunaeesti murderühmas seda aga kindlasti ei ole.

Põhjaeesti murderühma kuuluva keskmurde põhjal arvutati lõpuks z- ja t-skoorid sõnapaaridele „see+kes” ja „see+mis”. Eesmärgiks oli leida, kas sõnadevaheline seos on esimese sõnapaari puhul tugevam kui teise puhul, arvestades, et keskmurdes kasutatakse *kes*-i ka elututele sõnadele viitamisel. Tulemus näitas siiski, et seos on tugevam sõnade *mis* ja *see* vahel.

Tööga leiti küll eesti murretele iseloomulikke tunnuseid, kuid tõestati ka seda, et korrektset andmestikku selliste skriptidega ei saa. Andmed tuleb siiski ka lingvistil endal üle kontrollida.

Projektiga seotud dokumendid

Projektile on lisatud skriptide kasutamise juhend skriptide_kasutamine.txt, mis selgitab skriptide käivitamist, nende eesmärke, väljundeid ja kasutusjärjekorda. Ülejäänud projektile lisatud failid on siin loetletud vastavalt kaustadele, kus need paiknevad. Kui kaustas olevad failid on leitud skriptiga, siis on vastav skript toodud siin kausta failide nimekirjas esimesena. Kui kausta failidel on käivitatud järgmine skript, mis ei väljasta uusi faile (vaid näiteks hoopis failide ridade arvu), siis vastav skript on toodud siin kausta failide nimekirjas viimasena. GitHubi kaustades võib failide järjekord aga pisut erineda. Failide loetelu on koostatud samas järjekorras, milles neid ka töös endas kasutatakse.

Korpuse näitefailid:

- `AVI_Kaarel_Haav_EMH392_synt[UTF-8].xml` (IDAMURRE),
- `VMR_Leena_Kruusenber_synt[UTF-8].xml` (KESKMURRE),
- `KUL_Ants_Tammet_F0173_lihts[UTF-8].xml` (LÄÄNEMURRE),
- `ANS_Ann_Usin_F14_synt[UTF-8].xml` (SAARTE MURRE),
- `JOH_Liina_Laur_synt1[UTF-8].xml` (KIRDEMURRE),
- `HLJ_Marie_Pihlakas_EMH0135_synt[UTF-8].xml` (RANNAMURRE),
- `PST_Epp_Torm_EMH81_synt[UTF-8].xml` (MULGI MURRE),
- `KAM_Johannes_Pokk_EMH457[UTF-8].xml` (TARTU MURRE),
- `PLV_Leena_Kurvits_EMH_0222_synt[UTF-8].xml` (VÕRU MURRE),
- `ISE_Stepanida_Uibooss_F0311synt[UTF-8].xml` (SETU MURRE).

Kes_loendid:

- `kes_loend.sh`,
- `kes_loend_ida.txt`,
- `kes_loend_kesk.txt`,
- `kes_loend_laane.txt`,
- `kes_loend_saarte.txt`,
- `kes_loend_ranna.txt`,
- `kes_loend_kirde.txt`,
- `kes_loend_mulgi.txt`,
- `kes_loend_tartu.txt`,

- kes_loend_voru.txt,
- kes_loend_setu.txt.

Kes_sagedusloendid:

- kes_sagedusloend.sh,
- kes_sagedus_ida.txt,
- kes_sagedus_kesk.txt,
- kes_sagedus_laane.txt,
- kes_sagedus_saarte.txt,
- kes_sagedus_ranna.txt,
- kes_sagedus_kirde.txt,
- kes_sagedus_mulgi.txt,
- kes_sagedus_tartu.txt,
- kes_sagedus_voru.txt,
- kes_sagedus_setu.txt.

Kes_ühendid:

- kes_uhendid.sh,
- kes_uhendid_ida.txt,
- kes_uhendid_kesk.txt,
- kes_uhendid_laane.txt,
- kes_uhendid_saarte.txt,
- kes_uhendid_kirde.txt,
- kes_uhendid_ranna.txt,
- kes_uhendid_mulgi.txt,
- kes_uhendid_tartu.txt,
- kes_uhendid_voru.txt,
- kes_uhendid_setu.txt.

Kes_eelnev_sõna:

- kes_eelnev_sona.sh,
- kes_eelnev_sona_ida.txt,
- kes_eelnev_sona_kesk.txt,
- kes_eelnev_sona_laane.txt,
- kes_eelnev_sona_saarte.txt,
- kes_eelnev_sona_kirde.txt,
- kes_eelnev_sona_ranna.txt,
- kes_eelnev_sona_mulgi.txt,

- kes_eelnev_sona_tartu.txt,
- kes_eelnev_sona_voru.txt,
- kes_eelnev_sona_setu.txt.

Tähendus_mis:

- tahendus_mis.sh,
- tahendus_mis_ida.txt,
- tahendus_mis_kesk.txt,
- tahendus_mis_laane.txt,
- tahendus_mis_saarte.txt,
- tahendus_mis_kirde.txt,
- tahendus_mis_ranna.txt,
- tahendus_mis_mulgi.txt,
- tahendus_mis_tartu.txt,
- tahendus_mis_voru.txt,
- tahendus_mis_setu.txt,
- tahendus_mis_arv.sh.

Keskmurde skoorid:

- tekstid_koos_kesk.txt,
- korpusesuurus.sh,
- kes_arv.sh,
- see_arv.sh,
- kes_see_arv.sh,
- mis_arv.sh,
- mis_uhendid.sh,
- mis_see_arv.sh.

Kirjandus

Evert, Stefan 2007. Corpora and collocations. – Corpus linguistics. An international handbook 2. Toim. Anke Lüdeling, Merja Kytö. Berlin: Walter de Gruyter, 1212-1248.