doi:10.5128/ERYa11.17

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 11, 281–295

KEELETEADUSLIKE ANDMETE RUUMILISI VISUALISEERIMISVÕIMALUSI

Kristel Uiboaed, Aki-Juhani Kyröläinen

Ülevaade. Keeleteadusliku uurimismaterjali ja -tulemuste geograafiline visualiseerimine on dialektoloogias olnud alati kesksel kohal, kuid see pole omane ainult murdeuurimisele, vaid on oluline osa mis tahes lingvistilisest tööst, kuhu on kaasatud ruumiline komponent. Tänapäeval on olemas arvukalt programme ja kaardiressursse, mis võimaldavad neid ülesandeid täita. Sageli on aga nende kasutamise eelduseks üsna põhjalikud teadmised geoinfosüsteemidest või mõnest konkreetsest programmist. Käesoleva artikli eesmärk on pakkuda üks lihtne ja kiire võimalus geograafiliste andmete esitamiseks, eeldamata kasutajalt suuri tehnilisi oskusi. Pakutud lahendus ja materjalid on vabavarana saadaval ning iga soovija võib neid oma vajadustele vastavalt kohandada või selle baasilt välja töötada enda jaoks sobivad vahendid. Artiklis kasutame andmestikuna eesti murrete korpust, sh vadja keele alamkorpust, kuid võimalused eri andmete esitamiseks ei ole ühe konkreetse andmestikuga piiratud.*

Võtmesõnad: dialektoloogia, murdesüntaks, geolingvistika, korpuslingvistika, eesti keel

1. Sissejuhatus

Lingvistiliste nähtuste visualiseerimine kaartidel on alati olnud oluline osa murdeuurimisest. Traditsiooniliselt on selleks kasutatud olemasolevaid kaarte ning lingvistiline informatsioon on kaardile kantud enamasti käsitsi. Arvutiajastu on andmete esitusvõimalusi märkimisväärselt avardanud. Sellega on aga tehnoloogia keerulisemaks muutunud ja see eeldab lingvistidelt tunduvalt mitmekülgsemaid teadmisi ja tehnilisi oskusi. Selliste oskuste omandamine klassikalise keeleteadusliku hariduse juurde praegu ei kuulu. Artikli eesmärk ongi esitada üks võimalik viis geograafiliste andmete visualiseerimiseks, mis ei nõua spetsiifilisi tehnilisi oskusi ega pikemaajalist süvenemist eri programmidesse.

* Artikli valmimist on toetanud Eesti Teadusagentuuri projekt PUT90. Aitäh anonüümsetele retsensentidele parandusettepanekute eest ja Maarja-Liisa Pilvikule kommentaaride ja tagasiside eest.

Enamasti kasutatakse geograafiliste andmete visualiseerimiseks mitmesuguseid GIS-tarkvara (ingl *geographic information system*) võimalusi, mille rakendamise eelduseks on aga üsna põhjalikud teadmised GIS-süsteemidest ja geograafilisest terminoloogiast ning kindlad tehnilised oskused. GIS-süsteemid on sageli tasulised, mis seab nende kasutusele muidki piiranguid. Meie välja töötatud võimalus on lihtne viis andmete visualiseerimiseks. Kombineerime nii vabavaralist statistikaprogrammi R (R Development CoreTeam 2013) kui ka olemasolevaid vabavaralisi kaardiressursse, näiteks Google'i või Regio kaarte (2014). Siinne lähenemine ei eelda kasutajalt sügavamaid teadmisi geoinfosüsteemidest ega ka põhjalikke programmeerimisoskusi R-is (mõningad teadmised tulevad muidugi kasuks). Oleme artikli esitanud justkui kasutusjuhendi kõigi vajalike sammudega ning igal uurijal on juhiseid järgides võimalik oma andmestiku põhjal esitada soovitud infot. Eelduseks on vaid statistikaprogrammi R oma arvutisse paigaldamine. Käesoleva töö tulemusena valminud kaardiressursid, näidetes kasutatud failid ja R-i skriptid on kättesaadavad veebis ning kõik võivad neid vabalt kasutada ja vastavalt oma vajadustele muuta.[1] Viitamisjuhendid leiab veebilehelt ja failidest. Programmikoodi me artiklis põhjalikumalt ei kommenteeri, detailsem info on esitatud artikliga kaasnevas skriptifailis. Programmi R keeleteaduses rakendamise kohta on mitmeid väga häid õpikuid, mille abil on võimalik omandada sügavamaid teadmisi (Baayen 2008, Gries 2009, Johnson 2008). Ruumiliste andmete töötlemisest R-is on põhjaliku monograafia kirjutanud Bivand jt (2013).

Artikli alustuseks kirjeldame töö tausta ja põhjusi, miks me sellise lahenduseni oleme jõudnud. Kolmandas peatükis demonstreerime võimalusi, kuidas saada endale sobiv kaart, mida oleks võimalik kasutada edasiseks töötlemiseks ja oma uurimisandmete ja -tulemuste esitamiseks. Neljandas peatükis kirjeldame uurimisandmete, sagedusinfo ja statistilise analüüsi tulemuste kaartidel esitamist.

2. Taust

Praeguseks on olemas erinevaid visualiseerimis- ja analüüsirakendusi, mis on spetsiaalselt välja töötatud keeleteadlastele (nt Goebl 2006, Nerbonne jt 2011) lingvistiliste andmete analüüsimiseks ja uurimistulemuste esitamiseks. Need rakendused on hästi dokumenteeritud, nende kasutamine on tehtud lihtsaks ja mugavaks ka uurijatele, kes pole varem sarnaste töövahenditega kokku puutunud. Selliste rakenduste väljatöötamine ja olemasolu on väga tänuväärsed ning praeguseks võimaldavad need rakendused teha üsnagi keerulisi statistilisi analüüse. Nagu igasuguste valmistoodete puhul paratamatu, seavad nad kasutajale ka teatavaid piiranguid nii kaartide kasutamisel kui ka statistiliste meetodite valikul. Seega kui kasutaja soovib rakendada oma andmete analüüsimiseks meetodit, mida programmide arendajad pole rakendustesse kaasanud, pole kasutajal võimalik seda teha. Selleks tuleb otsida alternatiivseid lahendusi.

Tänapäeval on hulgaliselt kaardiressursse, mida saab kasutada ruumiliste andmetega seotud mis tahes töös, kaasa arvatud lingvistikas. Sellised kaardiressursid põhinevad sageli praegusel või ajaloolisel haldusterritoriaalsel jaotusel. Murdepiiride esitamiseks on selliste ressursside kasutamine aga mõnevõrra

[1] https://github.com/kristel-/Keeleteaduslike-andmete-ruumilisi-visualiseerimisvoimalusi (Uiboaed, Kyröläinen 2015).

problemaatiline. Traditsioonilised murdepiirid ei järgi enamasti haldusterritoriaalseid piire ja keelealad võivad ka riigipiire ületada. Selline problemaatika ei ole levinud ainult dialektoloogias. Näiteks on keeruline kasutada tänapäeva haldusterritoriaalseid kaarte ka juhul, kui uuritakse keelt, mis pole riigikeel ning pole seotud ühegi konkreetse tänapäevase või ajaloolise haldusüksusega. Näiteid pole vaja kaugelt otsida: probleem on tuttav kõigile väiksematele soome-ugri keelte uurijatele, rääkimata murdeuurijatest. Järgnevalt esitamegi ühe väga lihtsa võimaluse, kuidas selliseid kaarte ise joonistada, ning seejärel näitame, kuidas on võimalik kasutada oma andmete esitamiseks juba olemasolevaid elektroonilisi kaarte.

3. Kaardiandmestik

3.1. Poolkäsitsi joonistatud kaardid

Selles alapeatükis demonstreerime, kuidas üsna lihtsa vaevaga n-ö joonistada kaart, mida on võimalik automaatselt töödelda. Iga geograafilise andmestiku esitamise eeldus on koordinaatsüsteemil põhineva info olemasolu. Seega on meil alustamiseks vaja koordinaatide andmeid piirkonnast, mida soovime visualiseerida. Selleks on võimalik kasutada GIS-süsteemi ressursse. Suure tõenäosusega aga ei ole GIS-süsteemides kaarte, mis esitaksid nt liivi keele alad või eesti murdejaotuse. Niisiis on üks lahendus võtta mõni olemasolev kaart ja sellele toetudes n-ö joonistada uus kaart meid huvitavate piirkondadega. Keskne mõiste geograafias ja kartograafias on polügoon (ehk hulknurk), mis siinses kontekstis tähendab ühte kaardile kantud ala. Selleks on vaja teatavaid punkte, mida oleks võimalik ühendada, nii et neist moodustuks üks piirkond. Meie kasutasime Regio (2014) kaardisüsteemi, mis esitab pikkus- ja laiuskraadid. Sama eesmärgi täidab mis tahes sama infot esitav kaart. Erinevalt Google'ist pakub Regio infot ka kihelkondade kohta ning sobivad murdealad saamegi välja joonistada just kihelkonna piiride abil. Polügooni joonistamiseks piisab ainult mõnest punktist, kuid mida rohkem punkte, seda täpsem on kaart.[2]

Tabel 1. Fragment koordinaatandmestikust (failist *murdealadeKoordinaadid.csv*)

Murre	Latituud	Longituud
kirde	58.99	27.73
kirde	59.23	27.89
kirde	59.3	28.04
kirde	59.28	28.13
kirde	...	...
kirde	NA	NA

Tabel 1 esitab fragmendi failist *murdealadeKoordinaadid.csv* (Uiboaed, Kyröläinen 2015), mis sisaldab infot kõigi murdealade kohta ning nende alade joonistamiseks vajalikku infot. Iga murdeala (või ühe polügooni, näiteks saarte murdeala koosneb mitmest polügoonist) lõpetab rida, kus pikkus- ja laiuskraadide lahtrites on väärtus NA (ingl *not available*), mis on vajalik selleks, et polügooni joonistamine n-ö lõpetada või katkestada, et punkte ei ühendataks üle mitme polügooni. Oluline on

[2] Täpsemalt geograafilise terminoloogia ja vastava teoreetilise aluse kohta vaata Bibiko (2012) ja Krikmann (2005).

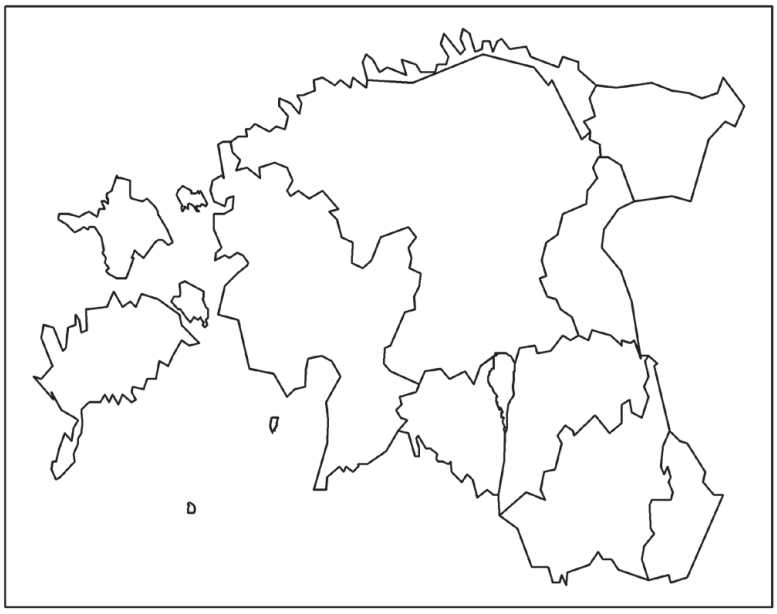
jälgida, et geograafilise andmestiku murdenimed vastaksid täpselt, ka kuju poolest, murdenimedele, mida kasutatakse lingvistilises andmestikus (täpsemalt vt allpool). Kõik tabelis esitatud koordinaadid on sisestatud käsitsi ehk kogutud punkt punkti haaval aluseks olevalt kaardilt (Regio 2014). Nagu eelnevalt mainitud, võib sama protsessi korrata mis tahes muul kaardiandmestikul, mis vajaliku info esitab.

Kui punktid on kogutud ja sobivas formaadis tabelisse kantud, on võimalik nende abil programmis R (R Development CoreTeam 2013) juba üsna lihtsalt kaart joonistada. Selleks tuleb andmestik esmalt R-i sisestada.

```
# loeme sisse kaardiandmestiku
koordinaadid <- read.csv("murdealadeKoordinaadid.csv", header=T,
sep=";")
```

Kui andmestik on programmi sisse loetud, piisab sellest lihtsakoelise kaardi joonistamiseks (tulemust vt kaardilt 1).

```
plot(koordinaadid$lon, koordinaadid$lat, type="l")
```



Kaart 1. Poolkäsitsi joonistatud eesti murdealade kaart

Järgnevas kirjeldame, kuidas sellele kaardile andmeid kanda ja seda näiteks sageduste visualiseerimiseks kasutada. Siin on oluline märkida, et parema tulemuse saab GIS-tarkvara[3] kasutades georefereerimise teel, mis on tehniliselt pisut nõudlikum viis, kuid tagab detailsema ja visuaalselt kvaliteetsema tulemuse. Täpsemalt vt selle kohta Kyröläinen ja Uiboaed (ilmumas).

[3] Vabavaraline GIS-tarkvara on näiteks Quantum GIS (QGIS Development Team 2014).

3.2. ggmap ja ggplot2 programmis R ning kasutatud kaardiressursid

Kui oma kaardi joonistamine pole hädavajalik ning sobivad olemasolevad kaardiressursid, on olukord veelgi lihtsam. R pakub selleks väga lihtsaid võimalusi, näiteks on muude võimaluste hulgas ka Google'i kaarte toetav pakett *ggmap* (Kahle, Wickham 2013), mida on üsna lihtne kasutada koos laialt levinud graafikapaketiga *ggplot2* (Wickham 2009). Alloleva skriptiosa käivitamisel saab edasiseks töötluseks alla laadida Eesti kaardi värvilise või mustvalgena (värvilise kaardi jaoks esitame vaid koodi). Tulemus on esitatud kaardil 2.

```
# vaikimisi värviline kaart
qmap("Eesti", zoom=7)

# mustvalge kaart
qmap("Eesti", zoom=7, color="bw")
```

Kaart 2. Paketiga *ggmap* saadud mustvalge Eesti kaart

Kaardi edasiseks töötluseks tuleks see salvestada järgneval viisil.

```
eesti <- qmap("Eesti", zoom=7, color = "bw")
```

Arvukate lisavõimaluste kohta (näiteks lisaks Google'i kaartidele on võimalik kasutada muidki kaarte) vaata paketi dokumentatsiooni (Kahle, Wickham 2013).

4. Keeleteaduslikud (meta)andmed ja nende esitus kaartidel

4.1. Sagedusandmete esitusvõimalusi

Eespool esitatud kaartidele on lihtne kanda sagedusandmestikku. Kasutame näitena finiitse verbi *saama* ja *tud*-partitsiibi konstruktsioonide andmestikku (Uiboaed 2013). Konstruktsiooniga saab eesti keeles väljendada nii passiivi kui ka impersonaali, selle kasutus varieerub murrete lõikes märkimisväärselt (Uiboaed 2013, Uiboaed jt 2013). Eespool esitatud kaardil võime visualiseerida selliste ühendite (mitte konstruktsioonide, st ühendi tähendus ei ole eristatud) sagedust niimoodi, et tumedamad alad tähistaksid kõrgemat sagedust ja heledamad vastupidi, madalamat sagedust ehk kasutame n-ö üleminekukaarti (ingl *gradience map*).[4] Sisendiks on tabel, kus on esitatud vaadeldava ühendi sagedus eesti murretes.

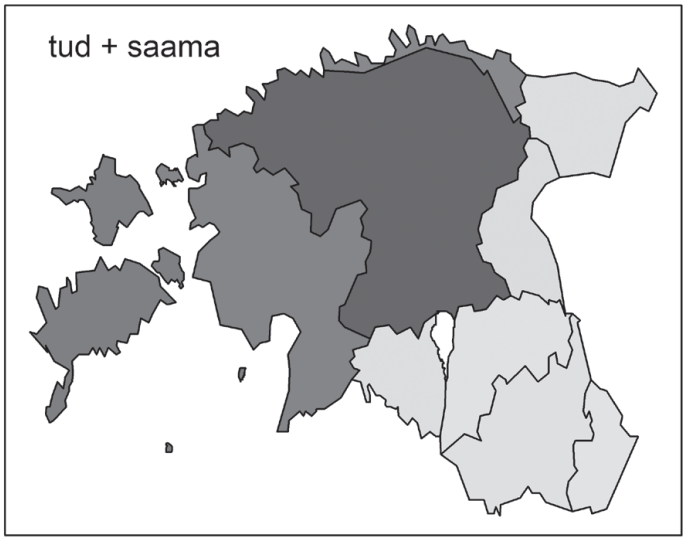
Tabel 2. Finiitse *saama*-verbi ja *tud*-partitsiibi ühendite sagedusandmestik (fail *tudSaama.csv*, Uiboaed, Kyröläinen 2015)

Murre	tudSaama
kirde	19
ida	29
tartu	13
võru	17
lääne	173
mulgi	21
seto	9
kesk	199
saarte	180
ranna	178

```
# kasutame juba tuttavat koordinaatandmestikku
koordinaadid <- read.csv("murdealadeKoordinaadid.csv", header=T,
sep=";")

# loeme sisse verbiühendite sagedusandmestiku
tudSaama <- read.csv("tudSaama.csv", header=T, sep=";")

# ja seejärel illustreerime neid sagedusi heledamate ja tumedamate
toonidega
plot(lat~lon, type="n", data=koordinaadid, axes=F, ylab="", xlab="")
for (i in 1: length(unique(koordinaadid$murre))) {
  temp=koordinaadid$murre==unique(koordinaadid$murre)[i]
  polygon(x=koordinaadid$lon[temp], y=koordinaadid$lat[temp],
     col=gray.colors(max(tudSaama$tudSaama))
[max(tudSaama$tudSaama)-tudSaama$tudSaama[tudSaama$murre==
unique(koordinaadid$murre)[i]]+1]
  )
}
box()
```


[4] Üleminekukaartide koodiosa on kirjutanud Pärtel Lippus.

```
text(x=min(koordinaadid$lon,na.rm=T), y=max(koordinaadid$lat,na.
rm=T)-0.1,labels="tud + saama",pos=4, cex=2.5)
```



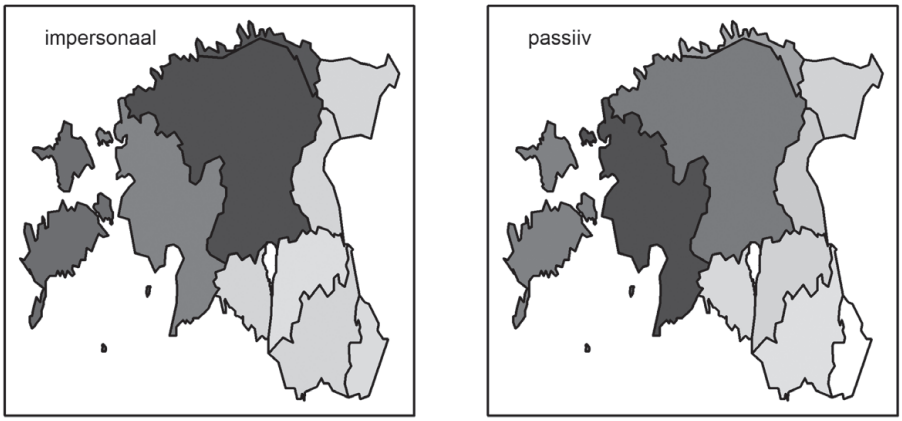
Kaart 3. Finiitse *saama*-verbi ja *tud*-partitsiibi ühendite kasutussagedus eesti murretes

Kaart 3 esitab finiitse *saama*-verbi ja *tud*-partitsiibi sageduste üleminekukaardi. Kaart illustreerib, kuidas selliste ühendite kasutussagedus esitatud andmestiku põhjal on suurim keskmurdes ja väiksem kõigis idapoolsetes murretes.

Samal viisil võime visualiseerida konstruktsioonide sagedust eraldi. Sagedusandmestik peab selle valiku korral olema jällegi eraldi tabelis (kasutame tabelit failis *tudKonstr.csv*, Uiboaed, Kyröläinen 2015).

```
# kasutame juba tuttavat koordinaatandmestikku (fail murdealadeKoor-
dinaadid.csv)
koordinaadid <- read.csv("murdealadeKoordinaadid.csv", header=T,
sep=";")

# loeme sisse konstruktsioonide andmestiku (fail tudKonstr.csv)
tudKonstr <- read.csv("tudKonstr.csv", header=T, sep=";")

# joonistame mõlema konstruktsiooni sageduste illustreerimiseks kaar-
did
plot(lat~lon, type="n", data=koordinaadid, axes=F, ylab="", xlab="")
for (i in 1: length(unique(koordinaadid$murre))) {
   temp=koordinaadid$murre==unique(koordinaadid$murre)[i]
   polygon(x=koordinaadid$lon[temp], y=koordinaadid$lat[temp],
     col=gray.colors(max(tudKonstr$imp))[max(tudKonstr$imp)-
tudKonstr$imp[tudKonstr$murre==unique(koordinaadid$murre)
[i]]+1]
   )
}
box()
text(x=min(koordinaadid$lon,na.rm=T), y=max(koordinaadid$lat,na.
rm=T)-0.1,labels="impersonaal",pos=4, cex=2.5)
```

```
plot(lat~lon, type="n", data=koordinaadid, axes=F, ylab="", xlab="")
for (i in 1: length(unique(koordinaadid$murre))) {
   temp=koordinaadid$murre==unique(koordinaadid$murre)[i]
   polygon(x=koordinaadid$lon[temp], y=koordinaadid$lat[temp],
     col=gray.colors(max(tudKonstr$pass))[max(tudKonstr$pass)-
tudKonstr$pass[tudKonstr$murre==unique(koordinaadid$murre)
[i]]+1]
   )
}
box()
text(x=min(koordinaadid$lon,na.rm=T), y=max(koordinaadid$lat,na.
rm=T)-0.1,labels="passiiv",pos=4, cex=2.5)
```



Kaart 4. Eesti murrete finiitse *saama*-verbi ja *tud*-partitsiibiga moodustatud impersonaali- ja passiivikonstruktsioonide sageduste üleminekukaardid

Kaart 4 esitab finiitse *saama*-verbi ja *tud*-partitsiibi abil moodustatud impersonaali- ja passiivikonstruktsioonide sagedused eesti murrete korpuse andmestiku põhjal, kus tumedamad alad tähistavad kõrgemat ja heledamad madalamad sagedust. Kaardilt 4 võib näha, et mõlema konstruktsiooni kasutussagedus on suurem läänepoolsetes murretes. Impersonaalikonstruktsioonid on levinumad keskmurdes ning passiivi kasutussagedus on suurim läänemurdes.

4.2. Geograafiliste punktide ja piirkondade esitamine

Sageli on vaja esitada ainult geograafilisi punkte, kust uurimismaterjali on kogutud, või märkida kaardil, millisest piirkonnast on mingit vormi, konstruktsiooni vms leitud. Selleks kasutatakse sageli nn sümbolkaarte, kus soovitud piirkonna võib kaardil tähistada lihtsalt punktiga. Eesti murrete kontekstis on klassikaliseks näiteks Andrus Saareste (1955) murdeatlase kaardid, mida oleks samuti võimalik esitada elektrooniliselt või kombineerida neid uute andmetega. Neid võimalusi me käesolevas artiklis ei esita, selle kohta vt täpsemalt Kyröläinen ja Uiboaed (ilmumas).

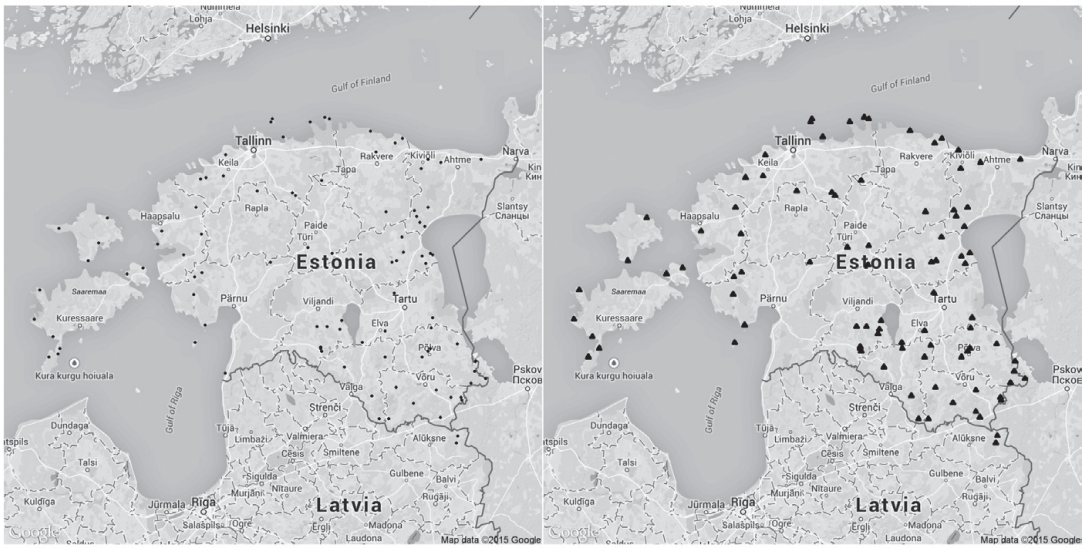
Sümbolkaartide esitamiseks on taas vaja tähistatavate piirkondade koordinaatandmestikku, mis on esitatud eraldi tabelina. Põhjana võib kasutada eespool esitatud kaarte.

Allolev kaart 5 illustreerib materjali ühest eesti murrete analüütilisi ja sünteetilisi kvotatiivikonstruktsioone käsitlevast uurimusest (Pilvik, Uiboaed 2014). Iga punkt kaardil tähistab küla, kust on töösse kaasatud uurimismaterjali kogutud. Näiteandmestikud on failides *kvotAn.csv* ja *kvotSyn.csv* (Uiboaed, Kyröläinen 2015).

```
# loeme sisse andmestikud failidest kvotAn.csv ja kvotSyn.csv, mis esitavad külad, kust tuvastati analüütilisi ja sünteetilisi vorme
kvotAn <- read.csv("kvotAn.csv", header=T, sep=";")
kvotSyn <- read.csv("kvotSyn.csv", header=T, sep=";")
# põhjana kasutame eelnevalt esitatud Google'i kaarti
# siin anname ette kaardi keskpunkti koordinaatidega, mitte kohanimega
eesti <- get_map(c(lon = 25.01361, lat = 58.59527), zoom=7, color = "bw")
```

Viimati nimetatud sammuga salvestame Eesti kaardi (objekt *eesti*), millele kanname andmestiku külade kohta, kust vaadeldavaid vorme tuvastati.

```
eesti_pohi <- ggmap(eesti)
eesti_pohi + geom_point(data = kvotAn, aes(x = lon, y = lat), size=1)
eesti_pohi + geom_point(data = kvotSyn, aes(x = lon, y = lat), size=1)
```



Kaart 5. Analüütiliste (vasakul) ja sünteetiliste (paremal) kvotatiivikonstruktsioonide tuvastamispunktid eesti murrete korpuse andmestikus (Pilvik, Uiboaed 2014)

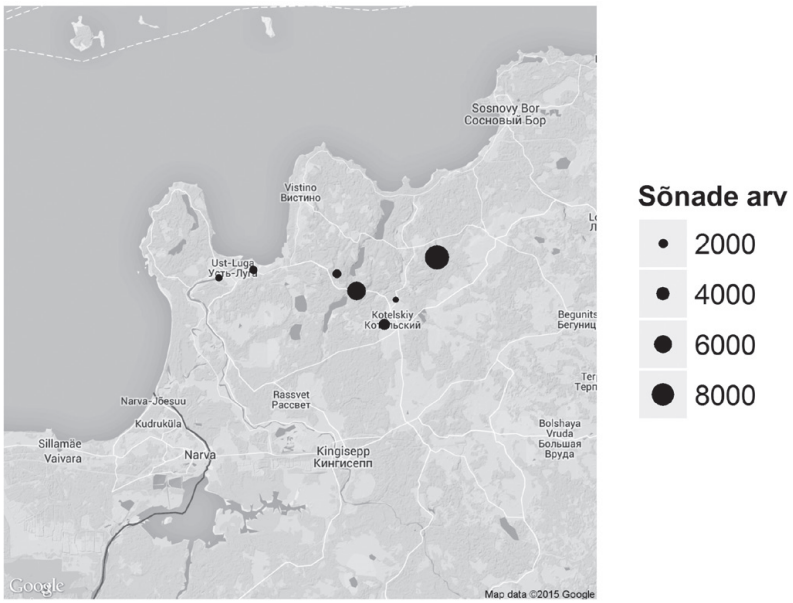
Iseenesestmõistetavalt ei piirdu esitatud andmete visualiseerimisvõimalused ainult eesti murdealadega. Sarnase kaardi võime visualiseerida ka vadja korpusandmestikuga. Eesti murrete korpuse (EMK 2014) üks osa sisaldab morfoloogiliselt märgendatud vadja keele tekste. Visuaalselt võime esitada sellise andmestiku analoogselt eelnevatega: tähistame kaardil geograafilised punktid, kust on korpusandmestiku tekstid pärit. Kuna Vadjamaad kui ametlikku haldusüksust ei eksisteeri, siis võib keskpunktiks võtta mis tahes sobiva piirkonna, olenevalt sellest, millist kaardikuju eelistada. Järgnevas näites kaardil 6 oleme selleks n-ö keskpunktiks valinud Habolovo järve.

```
# vadja keele alamkorpuse sagedusandmestik on failis vadja.csv, mille loeme järgneva käsuga R-i
vad <- read.csv("vadja.csv", header=T, sep=";", dec=",")

# laeme vajaliku kaardi
vadja <- get_map("Habolovo", zoom=9, color = "bw")
vadja_pohi <- ggmap(vadja,)

# esitame külad
vadja_pohi + geom_point(data = vad, aes(x = lon, y = lat), size=1)
```

Kaart 6. Murdekorpuse vadja keele alamkorpuse materjali kogumispunktid

Samale kaardile võime lisada rohkem visuaalset informatsiooni, näiteks muuta punktide suurust vastavalt sellele, kui palju materjali korpuses mingist külast sisaldub: punktide suurus on seotud sõnade arvuga konkreetse küla korpusandmestikus.

```
vad <- read.csv("vadja.csv", header=T, sep=";", dec=",")
vadja <- get_map("Habolovo", zoom=9, color = "bw")
vadja_pohi <- ggmap(vadja)
vadja_pohi + geom_point(data = vad, aes(x = lon, y = lat, size=sonu))
+
   scale_size_continuous(range = c(1, 4), name="Sõnade arv")
```



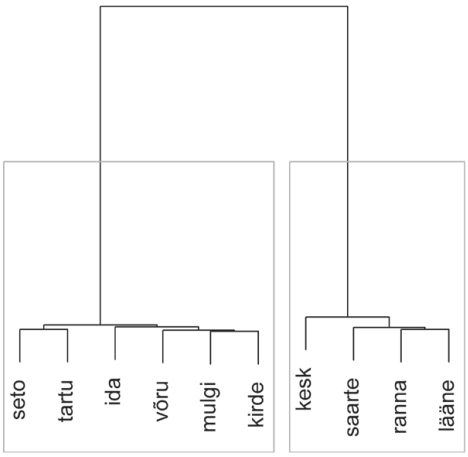
Kaart 7. Eesti murrete korpuse vadja keele alamkorpuse materjali kogumispunktid ja sõnade arv visuaalselt

Kaardil 7 on musta punktiga tähistatud koht, kust materjali on kogutud ning punkti suurus on viidud vastavusse sõnade arvuga igast punktist. Mida suurem punkt, seda rohkem on sellest külast korpuses sõnu ja vastupidi.

4.3. Statistilise analüüsi tulemuste esitamine

Järgnevalt näitame, kuidas samu vahendeid kasutades esitada kaardil klasteranalüüsi tulemusi. Klasteranalüüs on üks klassifitseerimismeetoditest, mis sarnaste tunnuste alusel grupeerib arvutuslikel põhimõtetel uuritavaid nähtusi (Everitt jt 2011). Siin me ei peatu meetodi üksikasjadel, vaid esitame lihtsalt ühe siinkirjutaja varasema uurimuse (Uiboaed 2013) klasteranalüüsi tulemused. Kasutame juba tuttavat *tud*-partitsiibi ja finiitse *saama*-verbi ühendite andmestikku. Klasteranalüüsiga saame dendrogrammi, mis on esitatud alloleval joonisel 8.

```
# analüüsiks loeme taas sisse andmestiku
klaster <- read.csv("tudSaama.csv", header=T, sep=";", row.names=1)

# arvutame kaugusmaatriksi, vajalik klasteranalüüsi sisendiks
andmed.kaugus <- dist(klaster, method="euclidean", upper=TRUE,
diag=TRUE)
klasteranalyys <- hclust(andmed.kaugus, method="ward.D")

# graafik
plot(klasteranalyys)
rect.hclust(klasteranalyys)
```



Joonis 8. Klasteranalüüsi dendrogramm *tud*-partitsiibi ja finiitse *saama*-verbi ühendite sageduse põhjal eesti murretes

Joonise 8 kohta võib lihtsustatult öelda, et sarnased objektid grupeeritakse samasse rühma ning moodustub kaks homogeenset ja teineteisest eristuvat rühma, mis on joonisel tähistatud heledamate joontega. Vasakpoolsesse kuuluvad Seto, Tartu, ida-, Võru, Mulgi ja kirdemurre ning teise, parempoolsesse rühma kesk-, saarte, ranna- ja läänemurre. Üsna sarnane pilt oli hoomatav ka sageduspõhistel üleminekukaartidel.

Järgnevalt kombineerime eelnevalt esitatud võimalusi ja esitame klasteranalüüsi tulemused ka geograafiliselt. Selleks kasutame faili *klasterTulemus.csv* (Uiboaed, Kyröläinen 2015), kus on esitatud polügoonide joonistamise järjekord, koordinaadid (sama andmestik, mis on failis *koordinaadid.csv*), murded ning klasteranalüüsi tulemusena saadud grupitunnus ja põhjana kasutame Google'i kaarti. Praegu visualiseerime vaid kaht analüüsis selgemalt eristunud gruppi.

```
klusterVis <- read.csv("klasterTulemus.csv", header=T, sep=";")

qmap("Eesti", zoom=7, color = "bw") +
geom_polygon(aes(x = lon, y = lat, group = murre, fill = klaster,
data=klusterVis))
```

Kaart 9. Klasteranalüüsi tulemused geograafiliselt

Kaart 9 esitab klasteranalüüsi tulemused kaardil, mis on kombineeritud ülal esitatud ise joonistatud kaardist ning Google'i kaardist. Sellisel viisil on kaks eelist. Ühelt poolt saame kasutada murdepiire, mis ei ühti haldusterritoriaalsete piiridega, teisalt muudame kaardi informatiivsemaks, lisades ühe lisakihina tänapäevase kaardi, mis annab võimaluse esitada uurimismaterjali ja -tulemusi kahes kontekstis.

5. Kokkuvõte

Artiklis esitasime ühe lihtsa lingvistiliste andmete kaartidel esitamise võimaluse, rakendades selleks statistikaprogrammi R ja olemasolevaid vabavaralisi kaardiressursse. Näitasime, kuidas on võimalik ise poolautomaatselt kaarti joonistada, kui olemasolevad kaardid pole piisavad, ning kuidas seda kaarti elektrooniliselt töödelda. Pakkusime välja võimalusi sagedusandmestiku esitamiseks kaartidel ning näitasime, kuidas võib kaartidel visualiseerida statistilise analüüsi, täpsemalt klasteranalüüsi tulemusi.

Artikli põhieesmärk oli pakkuda võimalikult lihtne lahendus kirjeldatud ülesannete täitmiseks, eeldamata kasutajalt sügavaid teadmisi geoinformaatikast ja eri programmidest. Välja töötatud ja rakendatud vahendid on vabalt kasutatavad (vt Uiboaed, Kyröläinen 2015) ning nende eeskujul on võimalik lihtsa vaevaga luua endale sobivad kaardid. Pakutud lahendus muudab lingvistiliste andmete ruumilise esitamise tehniliselt väga lihtsaks ja kiireks ning sobib suurepäraselt esmaseks analüüsi ja andmete visualiseerimiseks.

Viidatud kirjandus

Baayen, Harald R. 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.

Bibiko, Hans-Jörg 2012. Visualization and online presentation of linguistic data. – Frank Seifart, Geoffrey Haig, Nikolaus P. Himmelmann, Dagmar Jung, Anna Margetts, Paul Trilsbeek (Eds.). Potentials of Language Documentation: Methods, Analyses, and Utlization. Honolulu: University of Hawai'i Press, 96–104.

Bivand, Roger S.; Pebesma, Edzer J.; Gómez-Rubio, Virgilio 2013. Applied Spatial Data Analysis with R. New York–Heidelberg–Dordrecht–London: Springer.

EMK 2014. Eesti murrete korpus. http://www.murre.ut.ee/mkweb/ (1.12.2013).

Everitt, Brian S.; Landau, Sabine; Leese, Morven; Stahl, Daniel 2011. Cluster Analysis. 5th revised edition. Wiley Series in Probability and Statistics. Chichester: Wiley-Blackwell.

Goebl, Hans 2006. Recent Advances in Salzburg Dialectometry. – Literary and Linguistic Computing: Journal of the Association for Literary and Linguistic Computing, 21 (4), 411–435. http://dx.doi.org/10.1093/llc/fql042

Google Maps 2014.

Gries, Stefan Th. 2009. Statistics for Linguistics with R: A Practical Introduction. Berlin: Mouton de Gruyter. http://dx.doi.org/10.1515/9783110216042

Johnson, Keith 2008. Quantitative Methods in Linguistics. Malden: Blackwell Publishing Ltd.

Kahle, David; Wickham, Hadley 2013. ggmap: A package for spatial visualization with Google Maps and OpenStreetMap. R package version 2.3. http://CRAN.R-project.org/package=ggmap (15.2.2015).

Krikmann, Arvo 2005. Mart Remmeli "geograafilised figuurid". ["Geographical figures" by Mart Remmel.] – Joel Sang (Toim.). Endspiel. Kummardus Mart Remmelile. Tallinn: Eesti Keele Sihtasutus, 91–112.

Kyröläinen, Aki-Juhani; Uiboaed, Kristel (ilmumas). Visualization Techniques for Spatial Linguistic Data Using ggplot2 and ggmap.
Nerbonne, John; Colen, Rinke; Gooskens, Charlotte; Kleiweg, Peter; Leinonen, Therese 2011. Gabmap – A Web Application for Dialectology. – Dialectologia, Special Issue II, 65–89.
Pilvik, Maarja-Liisa; Uiboaed, Kristel 2014. Grammatical evidentiality in Estonian: Areal evidence of variation. – Ettekanne konverentsil Syntax of the World's Languages VI (SWL6), Pavia.
QGIS Development Team 2014. QGIS Geographic Information System. Open Source Geospatial Foundation. http://qgis.osgeo.org.
R Development CoreTeam 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing. http://www.R-project.org/.
Regio 2014. Eesti kihelkondade kaart. Regio AS. http://kaart.delfi.ee/.
Saareste, Andrus 1955. Petit atlas des parlers estoniens. Väike eesti murdeatlas. Uppsala: Almqvist & Wiksell.
Uiboaed, Kristel 2013. Verbiühendid eesti murretes. [Verb Constructions in Estonian Dialects.] Dissertationes philologiae estonicae Universitatis Tartuensis 34. Tartu: Tartu Ülikooli Kirjastus. http://hdl.handle.net/10062/34499
Uiboaed, Kristel; Hasselblatt, Cornelius; Lindström, Liina; Muischnek, Kadri; Nerbonne, John 2013. Variation of verbal constructions in Estonian dialects. – Literary & Linguistic Computing, 28 (1), 42–62.
Uiboaed, Kristel; Kyröläinen, Aki-Juhani 2015. Keeleteaduslike-andmete-ruumilisi-visualiseerimisvoimalusi. [Techniques for spatial data visualization in linguistics.] AndmestikeFailid. https://github.com/kristel-/Keeleteaduslike-andmete-ruumilisi-visualiseerimisvoimalusi (22.2.2015).
Wickham, Hadley 2009. ggplot2: Elegant Graphics for Data Analysis. New York: Springer.

Kristel Uiboaed (Tartu Ülikool) on eesti murrete teadur. Peamised uurimisvaldkonnad on keeleteaduslike andmete kvantitatiivne analüüs (eriti varieerumise uurimisel), korpuslingvistika, eesti murrete süntaktiline varieerumine, geograafiliste andmete esitusvõimalused.
Tartu Ülikool, eesti ja üldkeeleteaduse instituut, Jakobi 2, 51014 Tartu, Estonia
kristel.uiboaed@ut.ee

Aki-Juhani Kyröläinen (Turu Ülikool, Tartu Ülikool) on järeldoktor. Peamised uurimisvaldkonnad on statistilised ja katselised meetodid keeleteaduses ning uurimistulemuste ja -andmete visualiseerimine.
Tartu Ülikool, eesti ja üldkeeleteaduse instituut, Jakobi 2, 51014 Tartu, Estonia
akkyro@gmail.com

TECHNIQUES FOR SPATIAL DATA VISUALIZATION IN LINGUISTICS

Kristel Uiboaed[1], **Aki-Juhani Kyröläinen**[2,1]
University of Tartu[1], University of Turku[2]

Data visualization is an integral part of scientific inquiry in order to represent data and communicate findings. Recent developments such as the rise of large-scale corpora show that techniques to relate linguistically informed analysis and spatial data visualization have become increasingly important for quantitative analysis. Although spatial data visualization has gained momentum, these techniques may not be readily available for small or understudied languages. Here, we give an introduction to spatial data visualization using publicly available resources.

We use case studies on Estonian and Votic data to illustrate certain basic tasks in quantitative dialectology. We give solutions to create spatial maps based on either self-extracted coordinates or Google Maps. These maps can be used as a base layer and additional information, such as metadata and frequency distributions, can be represented on top of them. This approach can also be applied to statistical analysis. We illustrate this by carrying out a cluster analysis and its visualization using Google Maps. Thus, a toolkit is provided for quantitative analysis and spatial visualization in dialectology.

Keywords: dialectology, dialect syntax, geolinguistics, corpus linguistics, Estonian