KORPUSPÕHINE KVANTITATIIVNE DIALEKTOLOOGIA

LIINA LINDSTRÖM, MAARJA-LIISA PILVIK

Teaduse arenguloos on kirjeldatud mitmeid etappe, mis põhinevad oma ajas olemas olnud teadmistel ja oskustel ning nende põhjal tekkinud uuendustel. Nii on näiteks Thomas Kuhn, füüsik ja teaduse ajaloo uurija, eristanud 1962. aastal kolme suurt teaduse paradigmat või staadiumi, mis on seotud teaduse üldise kulu ning olemasolevate teadmiste kasutamisega. Neid staadiume võib pidada arenguetappideks, kuivõrd nad seostuvad mingi ajaperioodiga: 1) eksperimentaalne (empiiriline) teadus, mis kirjeldab reaalse maailma nähtusi ja protsesse ning loob esmase teadmiste baasi, seostub eelkõige renessansieelse teadusega, mil tehti enamik nendest avastustest, mida tänapäeval peame kooliharidus es iseenesestmõistetavateks baasteadmisteks; 2) teoreetiline teadus, mis modelleerib ja üldistab olemasolevaid teadmisi ning loob selle põhjal teooriaid, oli eriti aktuaalne arvutieelsel ajastul; 3) arvutuslik teadus, mis modelleerib ja simuleerib komplekssseid nähtusi arvutil, on olnud väga oluline suund kuni suurandmete kasutuselevõtuni (vt nt Kitchin 2014: 129). Sellele on lisatud neljas paradigma, mis on tänapäeval valdav: eksploratiivne teadus, mis põhineb suurtel andmehulkadel, modelleerib ja ühendab eri tüüpi andmeid (Hey jt 2009). Muutus andmekeskse lähenemise suunas on tänapäeva teaduses sedavõrd suur, et räägitakse ka andmerevolutsioonist (Kitchin 2014).

Humanitaarteadused, mis on klassikaliselt otsinud pigem inimkultuuriga seotud nähtuste t ä h e n d u s t ja mitte üritanud seda m õ õ t a (Hoyningen-Huene 2018), ei ole oma uurimisobjekti üldjuhul nimetanud a n d m e t e k s. Seoses digitaalsete vahendite ja uute meetodite plahvatusliku kasvuga on siiski ka humanitaarteadustes üha enam seda, mida on võimalik koguda, mõõta ja kokku lugeda, ning selle põhjal saada uuritavast nähtusest hoopis teistsugune pilt. Seetõttu räägime ka humanitaarteadustes üha enam andmetest. Andmekesksus ning meetodid nende andmete töötlemiseks ja analüüsiks on üks olulisemaid parameetreid, mille põhjal digihumanitaaria eristub klassikalisest humanitaariast.

Keeleteadus on läbi teinud teaduse üldised arenguastmed, võib-olla muidugi mitte päris samal viisil kui loodusteadused, mis olid Kuhni mudeli keskmes. Empiirilisel perioodil, mis kestis *ca* XIX sajandi keskpaigani, kirjeldati keeli ja nende struktuuri. Meetod tekkis keeleteadusesse esmakordselt seoses võrdlev-ajaloolise meetodi sünniga, kui hakati süstemaatiliselt võrdlema sugulaskeeli ning neis toimunud häälikumuutusi ja modelleeriti (rekonstrueeriti) nende põhjal keel(t)e varasemad arenguetapid ning algkeel. Seda võib pidada ka teooria sünniks. Teooriakesksed olid ka strukturalism ja hiljem generatiivne grammatika, viimasele vastureaktsioonina tekkinud

kognitiivne keeleteadus ja paljud muud lähenemised. Tänapäeval on keeleteadus aga oluliselt kvantitatiivsem ning andmekesksem kui muud humanitaarteadused; selle põhjuseks on tõenäoliselt arvutilingvistika kui arvutiteaduse ja keeleteaduse hübriidvaldkonna esiletõus ja areng alates 1960.–1970. aastatest ning keelekorpuste loomine. Korpuspõhiste ja kvantitatiivsete meetodite kasutamine on saanud tänapäeval nii üldiseks, et räägitakse kvantitatiivsest pöördest (vt Janda 2013; Levshina 2015). Kuivõrd muutus on seotud eelkõige suurte andmete käsitlemise ja analüüsiga, siis võiks seda nimetada ka andmepõhise keeleteaduse õitsele puhkemiseks. Andmepõhisel keeleteadusel on palju pioneere ja eestkõnelejaid, kes on koostanud ka vastavaid õpikuid, näiteks R. Harald Baayen (2008), Stefan Th. Gries (2009), noorematest Natalia Levshina (2015).

Kui küsida, kuhu selles pildis paigutub murrete uurimine, siis huvitaval kombel võib vastata, et murdeuurimine on olnud alati andmetel põhinev. Välitööd ning murdekeele dokumenteerimine on olnud vähemalt eesti dialektoloogias väga olulisel kohal alates 1920-ndatest, mil Emakeele Selts käivitas toonase Tartu Ülikooli õppejõu Lauri Kettuneni eestvõttel eesti murrete kogumise aktsiooni (Rätsep 2003). Need materjalid, millele hiljem on tulnud väga palju lisa, on tänini hoiul Eesti Keele Instituudi arhiivis ning Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivis. Küll aga on suuri koolkondlikke erinevusi selles osas, mida nende kogutud andmetega peale on hakatud või kuidas neid andmeid analüüsitud. Järgnev püüab anda ülevaate meetoditest ning suundadest dialektoloogias nii Eestis kui ka mujal. Seejuures on põhirõhk andmetega ümberkäimisel ning sellel, mida uut on kvantitatiivsed meetodid dialektoloogiale pakkunud.

1. Võrdlev-ajalooline paradigma ja murdeuurimise algus Eestis

Eesti murdeuurimise traditsioonid on välja kasvanud võrdlev-ajaloolisest meetodist, mille põhieesmärk on võrrelda sugulaskeeli ja murdeid, leida neis toimunud regulaarsed häälikumuutused ning selle põhjal rekonstrueerida keele varasem etapp ehk algkeel (vt nt ülevaadet Rankin 2003). Võrdlev-ajaloolise meetodi tähelepanu oli eelkõige vormi rekonstrueerimisel ning lähtuti häälikumuutuste seaduspärasustest. Keele sünkrooniline kasutus oli teisejärguline ning sellele läheneti pigem muutuste uurimise perspektiivist. Samamoodi läheneti ka murretele: kui vaatame XX sajandil Eestis koostatud klassikalisi murdeülevaateid, leiame sealt peatükid „Vokalism", „Konsonantism", „Noomen", „Verb", mis viitavad vastavalt vokaalide ja konsonantidega toimunud häälikumuutustele ning noomeni ja verbi muutmisega seotud detailidele ning muuteparadigmadele. Selle heaks näiteks on Aili Univere kirjutatud idamurde ülevaade (1988), aga sama skeemi järgivad teisedki samalaadsed tööd. Murrete kohta on koostatud ka detailsemaid üksikkäsitlusi, mis puudutavad vastavaid alaosi, näiteks Salme Nigoli kandidaaditöö Hargla murraku konsonantismi kohta, mis anti hiljem välja ka raamatuna (Nigol 1994). Eesti Keele Instituudi eesti murrete ja soome-ugri keelte arhiivis (EMSUKA) ning Tartu Ülikooli eesti murrete ja sugulaskeelte arhiivis (EMSA) hoitavad käsikirjalised murdeülevaated on reeglina koostatud samadel printsiipidel. Süntaksini

eesti murrete uurimise kuldajastul XX sajandil reeglina ei jõutud, vaid üksikutes töödes leiame märkmeid süntaktiliste nähtuste kohta (nt Must 1987; Juhkam 2012). Erandina väärivad märkimist Rein Nurkse põhjalik töö adjektiiviühildumise kohta (Nurkse 1937) ning Helmi Neetari käsitlused aluse ja öeldise ühildumise kohta eesti murretes (Neetar 1964, 1965a, 1965b).

Võrdlev-ajaloolise meetodi domineerimine eesti murdeuurimises on jätnud jälje üldisemalt sellele, kuidas murdeid kirjeldatakse või neist räägitakse. Nii näiteks on murrete kirjeldamises traditsiooniks lähtuda muutusest võrreldes varasema keelevormiga, ilma selgitamata, missugune see varasem vorm on. Nii leiame murdekirjeldustest sageli lauseid, nagu järgmises läänemurde ülevaates: „Enamikus Läänemaa kihelkondades on sõnaalgulisele *j*-ile järgnev *a* sõnuti asendunud *ä*-ga (*jänu*, *jäBur*, *jänDama*, *jätk*, *jäurama* jt.)" või „Põhja-Eestis, sealhulgas Läänemaal on toimunud häälikumuutus *er* > *är* (*pärà*, *pärast*, *äralDi*)" (Juhkam, Sepp 2000: 16). Implitsiitselt viidatakse neis kirjeldustes muutusele võrdluses algkeelega või keele varasema etapiga ning kirjelduse mõistmiseks peab seda teadma. Probleemiks võib taoline esitus olla lugejale, kes keeleajalugu võrdlev-ajaloolises raamistikus õppinud ei ole, sest tänapäeva lugeja jaoks oleks loogilisem võrdlusalus, mille suhtes keelt võrreldakse, tänapäeva standardkeel, mitte rekonstrueeritud algkeel.

Võrdlev-ajalooline meetod vajas algkeele rekonstrueerimiseks ohtralt andmeid eri murretest ja murrakutest ning selleks tehti süstemaatilist kogumistööd. Tänu sellele on meil tänapäeval kasutada väga palju murdeülevaateid, aga ka murdetekstide kirjapanekuid ja salvestusi, mida on võimalik kasutada ka teistsugustel eesmärkidel ning teisi uurimisküsimusi silmas pidades. Kuna murrete kogumist, talletamist ja uurimist peeti ka nõukogude perioodil väga oluliseks, on meie õnneks eesti keele kohta olemas väga suured kollektsioonid juba nimetatud Eesti Keele Instituudi ja Tartu Ülikooli arhiivides. Nende kahe kogu baasil on koostatud eesti murrete korpus (EMK), mis võimaldab uurida ka nähtusi, mis jäävad välja võrdlev-ajaloolise meetodi huvidest, nt jälgida keele nivelleerumist või teha kvantitatiivseid korpuspõhiseid uurimusi.

2. Atlastest dialektomeetriani

XX sajandi esimesel poolel levis Euroopas üha enam keelegeograafiline meetod, mille eesmärk oli koostada murdeatlasi. Murdeatlas esitab keele geograafilist varieerumist kaartidel. Eestis esindas seda suunda Andrus Saareste „Eesti murdeatlas" (I osa 1938, II osa 1941), „Väike eesti murdeatlas" (1955) ja tema suur käsikirjaline murdekaartide kogu Uppsala ülikooli arhiivis (RuRaKe).

Murdeatlaste koostamine eeldas hoopis teistsuguste andmete olemasolu: kaartidel saab esitada võrreldavaid andmeid, st sama nähtus peab olema fikseeritud laialt alalt erinevatest punktidest. Selline andmekogumine on nõudnud pikaajalist süstemaatilist tööd ning hulgaliselt kogumispunkte üle kogu vaadeldava ala. Saareste on „Väikeses eesti murdeatlases" välja toonud, et ta kasutas umbes 1600 küsimusest koosnenud küsitluskava ning andmed pärinesid umbes 500–550 küsitluskohast, millest kaartidele on jõudnud kuni 250 (Saareste 1955: 6). Saareste atlastes võib leida peamiselt (sõna)vormide ja sõnade kaarte, vähesel määral on ka muud. Seega põhineb iga kaart väga

suurel andmekogul, mis ei ole oma väärtust kaotanud tänapäevalgi. Osa kaartidest, sh Uppsala ülikoolis hoitavatest käsikirjalistest kaartidest on digiteeritud (RuRaKe).

Keeleatlasi (murdeatlasi) kasutatakse ka tänapäeval. Ühelt poolt pakuvad vanad atlased, millega kaasnevad hiiglaslikud andmekogud, palju võimalusi uuteks uurimusteks, millele omakorda annab tuge geoinfosüsteemide (GIS) areng: vanu kaarte saab digitaliseerida ja georefereerida (st siduda koordinaatidega ja seeläbi näiteks mõne teise kaardiga ning täpsemalt mõõta geograafilisi kaugusi punktide vahel), siduda uute või teist tüüpi (meta)andmetega, aga ka täiendada näiteks audio- või videomaterjalidega; samuti võib teha spetsiifilisi statistilisi analüüse, mis võtavad arvesse ruumiandmeid jne (Kretzschmar 2018). Näiteks Lauri Kettuneni koostatud Soome murdekaartide põhjal on tehtud katse modelleerida murrete kujunemist populatsioonigeneetikas kasutatavate meetoditega, mis võimaldavad hinnata murrete omavahelist sarnasust või erinevust (Syrjänen jt 2016). Terhi Honkola on oma väitekirjas sidunud sama soome murrete andmestiku teist tüüpi andmetega: kultuuriliste, keskkonna- ja administratiivandmetega ning geograafilise kaugusega, ning leidnud, et murretevahelised keelelised erinevused on tihedamalt seotud keskkonna- ja kultuuriliste mõjuritega kui geograafilise kaugusega (Honkola 2016: 155).

Murdeatlased on dialektomeetriliste uurimuste oluliseks sisendiks (vt nt ülevaadet Goebl 2018). Dialektomeetrilise meetodiga võrreldakse paljude üksiknähtuste alusel murrete omavahelisi keelelisi sarnasusi ja keelelist kaugust ning vaadeldakse üksikute murdejoonte asemel pigem üldist pilti, piltlikult öeldes puude asemel metsa (Nerbonne, Kleiweg 2007; Nerbonne, Kretzschmar 2013). Dialektomeetria teerajajaks peetakse Jean Séguy'd, täpsemalt tema uurimust gaskooni murretest (Séguy 1973). Arvutiajastul on valdkonda arendanud väga tugevalt Hans Goebl (vt ülevaadet Goebl 2018), hiljem ka John Nerbonne, William A. Kretzschmar, Jack Grieve jt. Dialektomeetria eeldab erinevate nähtuste koosvaatlust ning kvantitatiivsete meetodite rakendamist. Dialektomeetrilisi uurimusi võib läbi viia mis tahes keeletasandil (foneetika, fonoloogia, morfoloogia, süntaks, leksika, pragmaatika, semantika), kui vaid on olemas piisaval hulgal andmeid. (Nerbonne, Kleiweg 2007) Ka metoodika, kuidas keelelist kaugust mõõta, on mitmekesine ja võib lähtuda nii sarnasustest kui ka erinevustest (kauguse mõõtmise ja ruumiandmetele rakendatavate statistiliste meetodite kohta vt Grieve 2018; komplekssemate dialektomeetria meetodite kohta vt nt Grieve 2014; Nerbonne, Wieling 2018; Heeringa, Prokić 2018).

Eesti esimeste dialektomeetriliste uurimuste hulka kuulub Sirje Murumetsa kaheosaline artikkel, mis mõõdab eesti murrete sarnasusi ja erinevusi „Väikese murdesõnastiku" levikuandmete põhjal (Murumets 1982–1983). See põhineb suuresti Goebli väljatöötatud meetodil ning mõõdab Eesti kihelkondade vahelisi sõnavarasuhteid ja sõnavara ühisosa. Samas vaimus vaadeldakse sõnavarasuhteid ka Arvo Krikmanni ja Karl Pajusalu artiklis „Kus on keskmurde keskpunkt" (2000), mis põhineb samuti „Väikese murdesõnastiku" andmetel.

3. Variatiivsuse analüüs: sotsiolingvistiline murdeuurimine

Läänemaailmas pöörati 1950-ndatel üha enam pilke linnamurrete uurimisele, ehkki esialgu pigem hariduslike vajaduste võtmes. Üha enam märgati sotsiaalsete tegurite mõju urbaansete murrete keelekasutusele ning selles toimuvatele muutustele. (Le Page 1997) Variatiivsuse analüüs meetodina saigi alguse sotsiaalteaduste uurimismeetodite rakendamisest keelelise varieerumise uurimises ning selle suuna juhtfiguuriks sai William Labov.

Sotsiolingvistilise variatiivsuse uurimises vaadeldakse peamiselt keeleväliste sotsiaalsete ning keelesiseste variaablite mõju uuritava keelenähtuse varieerumisele: mis mõjutab ühe või teise keelevariandi eelistamist? Oluline on analüüsiks sobiva materjali olemasolu. Labovist alates on sotsiolingvistilise variatiivsuse uurimuste aluseks olnud spontaanne kõne, kus varieerumine ilmneb kõige loomulikumal kujul. Meetodi esmaseks sammuks on välja selgitada, missugused üldse võivad olla ühe uuritava keelelise nähtuse variandid: kas x ja y on sama nähtuse variandid või siiski erinevad nähtused? Varieerumisest saame rääkida vaid siis, kui x ja y võivad vastastikku üksteist asendada ilma muutusteta tähenduses või kasutuskontekstis (näiteks Audru murrakus varieerub *b* ja *v* kasutamine sõna sees: *kibi ~ kivi*, *leba ~ leva* 'leiva'). Järgmise sammuna kirjeldatakse konteksti, milles varieerumine võib aset leida (näiteks teatud struktuuriga sõnades; teatud süntaktilises konstruktsioonis jne). Seejärel formuleeritakse hüpoteesid tegurite kohta, mis võiksid varieerumist mõjutada (vt meetodi kohta nt Walker 2013; Paolillo 2002). Meetodi järgmine oluline samm on võimalike tegurite operatsionaliseerimine (analüüsitavaks tegemine) ning andmete süstemaatiline kodeerimine sellele vastavalt. Alles seejärel on võimalik läbi viia statistiline analüüs.

Peamiseks varieerumise statistilise analüüsi vahendiks on alates 1970-ndatest olnud programm VARBRUL (Cedergren, Sankoff 1974), mis võimaldab mõõta ja kontrollida eri tegurite tõenäosuslikku kaalu varieerumise mudelis. Eesmärk on välja selgitada, mis tegurid mõjutavad uuritava tunnuse ühe või teise variandi eelistamist ning missugune on tegurite koosmõju. VARBRUL kasutab arvutusliku meetodina logistilist regressioonanalüüsi, mis on ka tänapäeval asendamatu varieerumise uurimise meetod (vt nt meetodite võrdlust Tagliamonte, Baayen 2012). VARBRUL ja selle uuemad versioonid (GoldVarb) on kasutusel ka tänapäeval. VARBRUL-i edu saladus on kindlasti selle suhteliselt hõlpsas kasutatavuses: see ei nõua suuri eelteadmisi programmeerimisest ja statistikast ning sobib seetõttu hästi ka algajatele (Tagliamonte 2013). Tänapäeval on suurte andmekogude lisandumise, andmeteaduse hüppelise leviku ning vabavaraliste programmide (nt R ja Python) kasutuselevõtu ja kiire arenguga seoses suurenenud ka statistiliste meetodite hulk, mida variatiivsuse analüüsis kasutatakse: näiteks on laialt levinud klassifitseerimispuud ja juhumetsad, mis sobivad hästi ka juhul, kui andmestikus on „auke" ja kui andmed on ebaühtlaselt jaotunud (Tagliamonte, Baayen 2012), ning segamudelid, mis aitavad analüüsil arvestada ka juhuslike teguritega (vt lähemalt 4. ptk).

Eestisse jõudis varieerumise analüüs 1990-ndatel Karl Pajusalu eestvedamisel. Selle meetodiga on uuritud eelkõige murrete nivelleerumisest tingitud varieerumist: kuidas varieerumine toimub ning mis seda mõjutab, kui tradit-

siooniline regionaalne murre on kontaktis eesti ühiskeelega ja/või mingi teise keelekujuga. Meetodit järgides on uuritud näiteks Võrumaal Sute külas *n*- ja *h*-inessiivi vaheldumist (Pajusalu jt 1999; Velsker 2000; Mets 2010), *tud*-partitsiibi tunnuse, *olema*-verbi 3. pöörde mitmusvormi ning partikli *ikka* varieerumist (Mets 2010). Kui need uurimused põhinevad kitsa piirkonna keelekasutuse sotsiolingvistilisel analüüsil ning selles toimuvate muutuste jälgimisel, siis Lindström jt (2009) vaatasid 1. isiku pronoomeni kasutuse analüüsis laiemat eesti murrete ala ning võtsid ühe varieerumist mõjutava tegurina arvesse ka kihelkonda. See lähenemine seostub tugevamini juba korpuspõhise dialektoloogia peamise eesmärgiga murdeid omavahel võrrelda.

4. Korpuspõhine dialektoloogia

Keeleteaduses on korpuste loomine ning korpuslingvistika olnud produktiivne suund juba alates 1960.–1970. aastatest, mil loodi esimesed suuremad inglise keele korpused Brown ja LOB (vt McEnery, Hardie 2013). Eestis on keelekorpusi loodud alates 1980-ndate lõpust, mil asuti Browni korpuse eeskujul koostama eesti kirjakeele korpust (Hennoste, Muischnek 2000).

Lisaks kirjakeelele on loodud ridamisi ka spetsiifilisemaid eesti keele korpusi, nt vana kirjakeele korpus, suulise kõne korpus, suulise kõne foneetiline korpus, eesti murrete korpus jne. Korpuse moodustavad sarnastel alustel valitud tekstid, mis on tehtud elektrooniliselt kättesaadavaks, süstemaatiliselt töödeldud ja esitatud ning varustatud vajaliku metainfoga. Korpuse eesmärk on eelkõige teha andmed uurijale kättesaadavaks.

Eesti murrete korpust (EMK) on koostatud alates 1998. aastast Eesti Keele Instituudi ja Tartu Ülikooli arhiivide materjali baasil. See esindab kõige paremini 1870.–1890. aastatel sündinute suulist keelekasutust, mida on süstemaatiliselt kogutud ja salvestatud alates 1950-ndate teisest poolest välitööde käigus, hiljem transkribeeritud ning korpuse koostamise käigus viidud elektroonilisele kujule, märgendatud morfoloogiliselt ja osalt ka süntaktiliselt ning tehtud veebis kättesaadavaks (EMK koostamispõhimõtete kohta jms vt Lindström 2001, 2015).

Eesti on heas seisus, et sellise kollektsiooni kogumine ning taolise korpuse tegemine on olnud võimalik. Arvestades eesti murrete kiiret nivelleerumist XX sajandil ei oleks ilma EKI ja TÜ arhiivis säilitatud süstemaatiliste materjalideta paljude murrete kohta võimalik enam esinduslikke andmeid saada. Kui vaatame ringi laias maailmas, siis murdekorpusi on praegusel hetkel üsna väheste keelte kohta ning olemasolevad on omakorda koostatud mõnevõrra erinevatel alustel ning eesmärkidel. Nimetada võiks inglise (nt FRED), hollandi (DynaSAND), Skandinaavia keelte (Johannessen jt 2009), gruusia (GDC), portugali (Carrilho 2010), soome (Lauseopin arkisto), itaalia (ASIt) ja saksa (nt ArchiMob, REDE) keele murrete korpusi. Siiski korpuste arendamine jätkub ning tõenäoliselt lisandub lähiaastatel süstematiseeritud korpusi ka teiste keelte murrete kohta. Murdekorpuste teke on ühelt poolt seotud vastavate kollektsioonide olemasolu ning tehniliste võimaluste kiire arenguga, ent teisalt ka suurenenud huviga murdesüntaksi vastu, mida on varasemate meetoditega olnud raske uurida.

Korpuspõhisel dialektoloogial on mitmeid eeliseid võrreldes klassikalise atlasepõhise või ka sõnastikupõhise lähenemisega.

1. Korpus põhineb loomulikul keelekasutusel, murdekorpused enamasti suulisel loomulikul keelekasutusel. Korpustest saadavad keeleandmed on seega autentsed ning usaldusväärsed; atlasandmed võivad olla kallutatud, sest neid on enamasti kogutud küsitluste abil ning need iseloomustavad seega pigem teadvustatud keelekasutust. Küsitlustel põhinevad andmestikud võivad olla mõjutatud küsitleja keelekasutusest või küsitluse läbiviimise keelest ning risk, et tulemused on sellest tugevalt mõjutatud, on suurem kui korpuste puhul (vt nt ülevaadet Carrilho 2010), ehkki päriselt välistada ei saa seda ka korpustes (näiteks juhul, kui küsitleja keeleline taust on kõneleja omast väga erinev).

2. Korpus on universaalne andmekogu, mille põhjal saab analüüsida keele kõiki tasandeid (foneetikat, morfoloogiat, süntaksit, sõnavara jne), ehkki praktikas sõltuvad analüüsivõimalused suuresti korpuse koostamis- ja märgenduspõhimõtetest.

3. Korpuspõhine dialektoloogia võimaldab arvesse võtta varieeruvust ning eri variantide sagedust. Kui atlasandmed on tavaliselt esitatud viisil „Murdes X esineb variant a, murdes Y variant b, murdes Z esinevad nii a kui ka b", siis korpuspõhine lähenemine võimaldab esitada eri variantide sagedusi täpsemalt: „murdes Z esineb variant a 6 korda, variant b 24 korda". Korpusandmed esindavad lingvistilist reaalsust seega täpsemalt kui atlasandmed, kuivõrd need ei ole kategoorilised. „Korpuspõhine dialektomeetria on sageduspõhine dialektomeetria selle kõige puhtamas vormis" (Szmrecsanyi 2014: 92).

4. Korpuspõhise lähenemise eeliseks on uuritava nähtuse vahetu keelelise ümbruse kättesaadavus, mis loob võimalused analüüsida nähtuse semantilisi, pragmaatilisi, tekstilingvistilisi jm omadusi.

5. Korpuspõhise dialektoloogia abil saab uurida ka selliseid keelelisi nähtusi, mis atlastest tavaliselt välja jäävad kas nähtuse keerukuse, uurijate huvide või üldisema teadusliku mõtte arengu muutuste tõttu. Nii näiteks ei leia me Saareste atlastest tavaliselt infot erinevate pragmaatiliste partiklite (nt *noh*, *no*, *jah*, *ikka* jne) kohta, sest need ei olnud sel ajal keeleuurijate huvi fookuses. Samal ajal võivad murded sagedasemate partiklite kasutamise osas küllalt palju erineda (Lindström jt 2001).

6. Korpus sisaldab reeglina metainfot informantide kohta: nende vanus või sünniaeg, sugu, haridustase, tegevusala jne. See metainfo võimaldab uurida varieerumist ka sotsiolingvistiliste parameetrite põhjal.

7. Korpus võimaldab uurida eelkõige sagedasi nähtusi, mis traditsioonilistes uurimustes (ja ka atlastes) jäävad sageli tähelepanuta. Näiteks varasemad murdealased tööd on keskendunud pigem neile joontele, mis murretes erinevad kirjakeelest, ja vähem neile, mis on ka kirjakeeles sagedased. Kvantitatiivne lähenemine annab võtme, kuidas sagedastele nähtustele läheneda. Näiteks on meie uurimisrühma töödest välja tulnud, et murrete võrdluses võib ka n-ö tavaline nähtus kvantitatiivses perspektiivis huvitavaid tulemusi anda (vt nt mineviku liitaegadega seonduvat Lindström jt 2015, 2018).

Mis laadi uurimusi on võimalik läbi viia murrete korpuse põhjal? Järgnevas ülevaates toetume osalt Benedikt Szmrecsanyi ja Lieselotte Anderwaldi (2018) artiklile ning teisalt eesti murrete korpuse põhjal tehtud töödele, milles

oleme olnud ise osalised või millega oleme muidu hästi kursis. Sellest tulenevalt on näited murrete grammatika alased ning kirjeldatud meetodid sobivad pigem loendusandmete ning kategoriaalsete tunnuste analüüsiks.

4.1. Kvalitatiivne analüüs

Murdekorpuse üks kasutusvõimalusi on üksiknäidete analüüs eesmärgiga osutada murrete varieerumise kvalitatiivsele küljele: milliseid variante üldse murdematerjalides esineb ning mille poolest need üksteisest või standardkeelest erinevad, missuguseid tähendusi/funktsioone kannavad jne. Korpus sobib selleks hästi seetõttu, et see on veebi kaudu hõlpsasti kättesaadav, analüüsitud ning mitmekesiste päringuvõimalustega. Eesti murrete korpuse veebipõhine otsimootor (MKWEB) võimaldab otsida nii sõne, lemma, sõnaliigi kui ka morfoloogilise info põhjal ühe tunnuse kaupa. Keerukamaid otsinguid, kus võetakse arvesse korraga mitu tunnust (näiteks ühe sõna lemma ja teise sõna grammatiline vorm), on võimalik sooritada XML-failide põhjal ning see nõuab mõningast skriptimisoskust.

Kvalitatiivse analüüsi eesmärgiks võib olla nii kasutusmustrite kirjeldamine kui ka võrdlemine muu allkeelega: nt *des*- ja *mata*-konverbitarindite võrdlus (Plado 2015), *mine*-teonimekonstruktsioonide kasutusvõimaluste kirjeldamine (Pilvik 2017), essiivi jäänukite analüüs võrdluses kirjakeelega ning muude keeltega (Metslang, Lindström 2017). Kvalitatiivset analüüsi võib kombineerida ka kvantitatiivsega (nt partitiivsubjekti kasutusalade kohta eesti murretes (Lindström 2017).

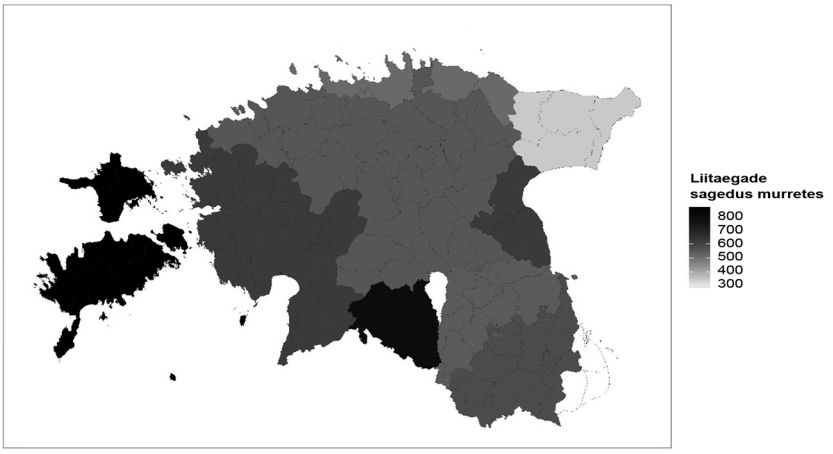
Ehkki võib tunduda teisiti, nõuab ka kvalitatiivne analüüs piisava suurusega andmehulka, et olla representatiivne. Seda saab loomulikult teha ka väikese andmehulga põhjal, ent tõenäoliselt ei anna see adekvaatset pilti murrete tegelikust keelelisest reaalsusest – oht on näiteid üle interpreteerida, seda eriti tekstides suhteliselt harva esinevate nähtuste korral; samuti on suur tõenäosus, et paljud murdele iseloomulikud kasutusviisid ei tule väiksesest andmestikust välja.

4.2. Kvantitatiivne ühe joone analüüs

Kvantitatiivne ühe joone analüüs võimaldab vaadelda ühe nähtuse levikut ning selle sageduserinevusi murretes ehk nn mikrotasandi varieerumist. Kvantitatiivne analüüs eeldab kvantitatiivsete meetodite rakendamist ning andmete visualiseerimist. Sellel on kaks põhilist eesmärki: andmete kirjeldamine ja andmete seletamine.

Kvantitatiivse analüüsi alla liigitub eraldi sagedusele orienteeritud analüüs, milles vaadeldakse eelkõige keelelise joone levikusagedust murretes. Näiteks eesti murrete põhjal on läbi viidud mineviku liitaegade esinemissageduse uurimus, mis ei vaatle mitte liitaegade vormistamise detaile, vaid esinemissagedust üldiselt. Kuna eesti keele liitajad on kujunenud pikka aega tagasi keelekontaktide kaudu (enamasti räägitakse balti ja germaani keelte mõjust minevikuaegade väljakujunemisel, vt Ariste 1956; Serebrenni-

kov 1959; Ikola 1960; Laanest 1975; Laakso 2001), siis võib oletada, et sageduserinevused võivad olla vähemalt osalt tingitud keelekontaktidest. See tuli ka meie liitaegade uurimusest välja: germaani keeltega kontaktis olnud aladel (saared, Lääne-Eesti) oli liitaegade kasutussagedus kõrgem kui aladel, mis on olnud kontaktis vene keelega või on olnud üldiselt konservatiivsemad, vähemalt germaani mõju suhtes (vt täpsemalt Lindström jt 2015, 2018). Samas ei olnud liitaegade seos teadaolevate keelekontaktidega sugugi alati ilmne: näiteks oli liitaegade sagedus kõrge ka Mulgi alal, ent seda on raske seostada konkreetse(te) keel(t)e mõjuga; põhjuseks võib olla ka muid tegureid, näiteks enneminevikku sage kasutamine evidentsiaalse strateegiana. Liitaegade kõrge kasutussagedus saarte murdes ning osalt ka läänemurdes võib aga lisaks kontakti mõjule olla seotud üldisemate muutustega eituse vormistamises, sest andmetest tuli selgelt esile täisminevikku eitavate vormide kõrge kasutussagedus koos sõnaga *pole*: *pole* + *nud*-konstruktsioon (*pole käinud*) on üle võtnud lihtminevikku eituse (vt lähemalt Lindström jt 2015, 2018). Liitaegade kasutussagedus on visualiseeritud joonisel 1.



Joonis 1. Täis- ja enneminevikku kasutussagedus eesti murrete korpuses (Lindström jt 2015).

Ühe joone sagedusele orienteeritud analüüsis peab arvestama, et on nähtusi, mille puhul erinevused on kas ebaolulised või raskesti seletatavad. Näiteks partitiivsubjekti kasutussageduse erinevused eesti murretes on raskesti seletatavad keelekontaktide mõjuga, nagu selgitas oma artiklis Liina Lindström (2017); seetõttu on seda kombineeritud kvalitatiivse analüüsiga, et leida pigem tüüpilisemad mustrid ja erinevad kasutusviisid ning esile tuua kirjakeele põhjal tehtud uurimustes puuduvat.

Peamiseks kvantitatiivseks meetodiks, mida sellistes sagedusele orienteeritud analüüsides rakendada, on sageduserinevuse olulisuse hindamine murrete vahel või võrreldes referentskorpusega (nt kirjakeelega), samuti muutustrendide väljatoomine (sageduse kasvamine või kahanemine kahe erinevast

ajast alamkorpuse võrdluses). Need meetodid eeldavad seega mingi võrdluseks sobiva teise (alam)korpuse või andmestiku olemasolu.

Kui vaadata sügavamalt keelematerjali sisse ning võtta arvesse keelenähtusega seotud eri kasutus- või tähendusfunktsioone, on võimalik kasutada ka meetodeid, mis visualiseerivad keelejoone iseloomulikke kasutusviise ning aitavad esile tuua seda, mis on mingitele murretele iseloomulikum, mis vähem iseloomulik, millised murded on selle keelejoone põhjal omavahel sarnasemad jne. Selleks võib kasutada näiteks korrespondentsanalüüsi (Greenacre 2007; Lebart jt 1998), mis sobib hästi just kategoriaalsete tunnuste vaheliste sagedu spõhiste seoste tuvastamiseks. Seda on rakendanud nt Kristel Uiboaed (2013) verbikonstruktsioonide analüüsis, Mirjam Ruutma (2016) pre- ja postpositsioonide kasutuse analüüsis. Murrete koondumist mingi(te) valitud tunnus(t)e põhjal saab kuvada ka klasteranalüüsi (Everitt jt 2011; eesti murretel rakendanud Uiboaed 2013) või mitmemõõtmelise skaleerimise abil (Meyers jt 2006). Kõik need meetodid põhinevad ühel või teisel moel (normaliseeritud) sagedustabelitel, mis peegeldavad tunnustevahelisi seoseid, ning võimaldavad esitada sarnasusi ja erinevusi, tunnustevahelisi tõmbumisi ja tõukumisi graafiliste kaugustena.

Piirangutele orienteeritud ühe joone analüüs keskendub murdejoone kahe (või enama) variandi varieerumist mõjutavate piirangute ja tegurite väljaselgitamisele. Sisuliselt on tegemist sama meetodiga nagu sotsiolingvistikas kasutatud klassikaline varieerumise analüüs (vt 3. ptk), mille puhul uuritakse võrdväärsete alternatiivide kasutamismustreid, ent korpuspõhises murdeuurimises on fookus sotsiaalsetelt parameetritelt ja ühe piirkonna keele varieerumise uurimiselt liikunud grammatika, st morfosüntaksi uurimisele ning piirkondlike variantide võrdlemisele. See omakorda tekitab vajaduse ka mahukamate andmestike kasutamiseks.

Võrdväärsed variandid tähendavad reeglina seda, et need võivad esineda samas positsioonis (nt *b~v* varieerumine: *kivi ~ kibi*) või olema funktsionaalselt või semantiliselt võrdväärsed (leksikaalsed variandid: *tarvis* või *vaja*; morfoloogilised variandid: *-n*, *-h* või *-s* inessiivi lõpuna, nt *külä-n ~ külä-h ~ küla-s*; süntaksis relatiivlause moodustamise variandid: *mees,* kes istus pingil ~ pingil istunud *mees*). Variatiivsuse analüüsi kaasatakse sõltumatuid/seletavaid tunnuseid, mille mõju sõltuva/uuritava tunnuse variantide esinemisele üritatakse analüüsi käigus välja selgitada. Murdekorpuse materjalidele meetodit rakendades võib ühe sõltumatu tunnusena arvesse võtta ka murret (või murrakut), et välja selgitada, kui suur roll on murdest tingitud erinevustel võrreldes muude keelesiseste ja -väliste tunnustega.

Vaid ühe sõltumatu/seletava tunnuse mõju uurimisel sõltuva/uuritava tunnuse varieerumisele võib kasutada kas Pearsoni (arvuliste tunnuste puhul) või Spearmani korrelatsioonikordajat (arvuliste või skalaarset hinnangut näitavate tunnuste puhul) või morfoloogia- ja süntaksiuurimustes sagedamini rakendatavat hii-ruut-statistikut, mis sobib kategoriaalsete tunnuste hindamiseks (nt Hanna Pook (2018) on kasutanud hii-ruut-testi, et võrrelda elutule objektile viitava asesõna *kes* kasutust murretes seoses erinevate lausetüüpide ning käänetega). Hii-ruut-statistik näitab, kas kahe tunnuse jaotus andmetes on juhuslik või on alust uskuda, et ühe tunnuse või tunnuse kategooriate proportsioonid erinevad oluliselt teise tunnuse (kategooriate) jaotuses.

Sageli aga ei huvita grammatikauurijaid pelgalt üksikute tunnuste omavahelised seosed, vaid ühe tunnuse varieerumine kontekstis, kus omakorda varieeruvad mitmed tegurid. Selleks et seletada kategoriaalse tunnuse varieerumist mitme teguri ning nende võimalike koosmõjude kaudu, on pikka aega kasutatud logistilist regressiooni, mis on ka 3. peatükis mainitud VARBRUL-i aluseks. Logistilise regressiooni mudeli põhiväljundid on koefitsiendid, mis peegeldavad keelelise nähtuse ühe variandi esinemise šansse võrreldes teise variandiga ja nende šansside muutumist konteksti muutudes (nt 1. isiku pronoomeni esinemise *vs.* mitteesinemise šansse vastavalt sellele, kas sellega koos esineval verbil on pöördelõpp, kui kaugele jääb diskursuses viimane 1. isiku viide, millise murdealaga on tegemist jne).

Üha enam on korpuspõhises keeleuurimises varieerumise seletamiseks hakatud kasutama segamudeleid ehk hindama lisaks fikseeritud seletavate tunnuste mõjule ka juhuslike tunnuste mõju regressioonimudelites. Fikseeritud tunnustel on kindel arv tasemeid ning iga tase pakub uurijale omaette huvi. Sellised tunnused on enamasti näiteks grammatilised kategooriad, nagu kääne, arv, isik, pööre, polaarsus jne. Juhuslike tunnuste puhul on andmetes esindatud alati juhuslik valim kõikidest võimalikest tunnuse tasemetest ning uurijale pakub huvi pigem tunnuse üldine mõju varieerumise seletamisele. Juhuslikud tunnused on korpuspõhises murdeuurimises näiteks kõnelejad, ent grammatiliste nähtuste uurimisel võib juhuslikeks tunnusteks pidada ka lekseemid, mida mingis grammatilises vormis realiseeritakse. Segamudelid aitavad seega arvestada näiteks tõsiasjaga, et mõned sõnad on korpustes sagedamad ja mõned kõnelejad räägivad rohkem. Murdealasid, mille jaotused on tekkinud eelkõige haldusüksuste põhjal (murrakud enamasti kattuvad kihelkondadega) ja millel puuduvad ranged üleminekud, võib siinjuures vaadelda nii fikseeritud kui ka juhuslike tunnustena.

Eelkõige viimasel kümnendil on keele varieerumise uurimustes regressiooni kõrval või suisa selle asemel lisaks teistele masinõppemeetoditele jõuliselt kanda kinnitanud klassifitseerimispuud (ingl *conditional inference tree*) ja juhumetsad (ingl *random forest*) ehk nn puude ja metsade mudelid (Breiman jt 1984; Strobl jt 2009). Nende meetodite suureks eeliseks on see, et need ei esita andmete adekvaatseks analüüsiks niivõrd suuri piiranguid andmete jaotusele kui näiteks regressioonimudelid, olles samal ajal meetoditena sama võimsad, meetodi rakendamisel kasutaja suhtes vähemnõudlikud ning väljundi poolest intuitiivsemalt tõlgendatavad. Puude ja metsade mudelid annavad informatsiooni nii uuritava tunnuse ennustatava väärtuse kohta mingis kontekstis kui ka mudelisse kaasatud tunnuste suhtelise olulisuse kohta uuritava tunnuse varieerumise seletamisel. Eesti murdeuurimises on klassifitseerimispuid rakendanud nt Ruutma jt (2016), analüüsides nii eeskui ka tagasõnadena esinevate kaassõnade paiknemist, Lindström ja Uiboaed (2017), analüüsides varieerumist *tarvis*/*vaja*-konstruktsioonides, ning Lindström jt (2018), analüüsides nimetamiskonstruktsioone (nt *Seda kutsutakse Hobusekiviks*). Metodoloogilise pluralismi näiteks murdeandmete kasutamisel võib aga pidada Jane Klavani jt (2015) uurimust kaassõna *peal* ja adessiivi varieerumisest, kus logistilise regressiooni segamudel ning puude ja metsade mudelid üksteist andmetes esinevate varieerumismustrite seletamisel hästi täiendasid, seletades varieerumise erinevaid aspekte.

Mõnevõrra uue suunana nn tõenäosuslike grammatikate kasutamisel murdeuurimises on kõikide keelevariantide (nt murrete) korraga mudeldamise asemel hakatud tähelepanu pöörama sellele, kui hästi seletavad mingid tegurid uuritavat nähtust murretes eraldi ja kuivõrd sarnased või erinevad on murded tegurite olulisuse osas. Sellist keelevariantide „oma grammatikate" võrdlust on tehtud nt inglise keele piirkondlike variantide põhjal (nt Grafmiller jt 2017; Szmrecsanyi jt 2016, 2017), ent meetodit sobiks rakendada ka väiksema piirkonna keelevariantide võrdluseks, samuti murrete ning nende kontaktkeelte konvergentsi ulatuse ja põhjuste tuvastamiseks.

4.3. Korpuspõhine dialektomeetria

Kvantitatiivne mitme joone analüüs ehk korpuspõhine dialektomeetria (ka: kobardialektoloogia, vt Uiboaed 2013) hõlmab ühtaegu paljusid keelelisi jooni ning nende varieerumist murretes. Erinevus klassikalisest dialektomeetriast seisneb selles, et andmed pärinevad korpusest ning see võimaldab arvesse võtta ka nähtuse/variantide esinemissagedust. Eristatakse ülalt-alla ja alt-üles korpuspõhist dialektomeetriat. Ülalt-alla dialektomeetria puhul on eelnevalt välja valitud analüüsi kaasatavad jooned, nende sagedused ja/või tõenäosused ning joonte põhjal arvutatud murrete lingvistilised kaugused. Alt-üles dialektomeetriline analüüs ei põhine mitte etteantud joontel, vaid korpusest mingil viisil ekstraheeritud andmetel, näiteks N-grammidel (bi- või trigrammidel) vms. (Wolk, Szmrecsanyi 2016; Szmrecsanyi, Anderwald 2018)

Korpuspõhise uurimuse hea näide on Benedikt Szmrecsanyi briti murrete süntaksi uurimus. See põhineb 57 morfosüntaktilise joone sagedusinfol, mis on saadud Freiburgi inglise murrete korpusest (FRED). Andmed on pärit 368 murdeintervjuust, mis on saadud 427 informandilt 158 geograafilisest punktist. Saadud sagedused on parema võrreldavuse huvides normaliseeritud. Analüüsi kaasatud joonte kohta on moodustatud N x p sagedusmaatriksid (milles N = 34 murret ridadena ja p = 57 joont tulpadena). Sagedusmaatriksi kaasatud 57 joone põhjal on seejärel arvutatud iga kahe murde vahelised eukleidilised kaugused. (Szmrecsanyi 2013) Analüüsi olemus on metodoloogiliselt sama nagu eelmises alapeatükis kirjeldatud sageduspõhise murrete klassifitseerimise puhul, ent siin on tegemist lihtsalt nähtuste kuhjamisega (ingl *aggregation*): kaht uuritavat objekti (murret) iseloomustavate tunnusekomplektide alusel on saadud mõõdikud, mis võimaldavad hinnata lingvistilisi kaugusi. (Eukleidilise kauguse arvutamise kohta eesti keeles vt nt Tooding 2015: 363–366.) Eukleidiliste kauguste põhjal on võimalik moodustada kauguste maatriks, mis iseloomustab iga kahe mõõtmispunkti (murde) vahelist lingvistilist kaugust. Seda on omakorda lihtne võrrelda geograafiliste kaugustega. Eukleidilised kaugused sobivad meetodina korpuspõhiseks analüüsiks hästi ka seetõttu, et need võtavad arvesse erinevate joonte sageduserinevusi: suuremad sageduserinevused saavad suurema kaalu kui väikesed sageduserinevused (vt meetodi kohta täpsemalt Szmrecsanyi 2013: 24–31).

Sagedusmaatriksit, mis oli eukleidiliste kauguste arvutamise aluseks, on võimalik kasutada ka muude statistiliste meetodite sisendina, mis samuti võimaldavad analüüsida murrete omavahelist (keelelist) lähedust või kaugust

suure hulga tunnuste põhjal. Sellised meetodid on näiteks peakomponentanalüüs (mida kasutas Szmrecsanyi 2013), klasteranalüüs ja korrespondentsanalüüs (vt ka ptk 4.2). Kobardialektoloogias (dialektomeetrias) kasutatavate statistiliste meetodite kohta vt lähemalt nt Grieve (2014) ning Nerbonne ja Wieling (2018).

Eestis on kobardialektoloogia meetodeid kasutanud Kristel Uiboaed (2013) oma doktoritöös eesti murrete verbiühendite uurimisel eesti murrete korpuse põhjal (vt ka Uiboaed jt 2013). Selles töös käsitles ta sagedasemate finiitsete verbide ühendeid koos kindlate infinitiivsete verbivormidega (nagu näiteks *sai teha*, *sai tehtud*, *käis söömas* jne) eesmärgiga välja selgitada, kas murretes on olulisi kasutuserinevusi. Verbiühendite analüüsil tuli välja, et Eesti lääne- ja põhjapoolsed murded kasutavad verbiühendeid enam kui lõuna- ja idapoolsed; samuti oli erinevusi eri ühendite kasutuses.

Kobardialektoloogiliste meetodite rakendamise peamiseks raskuspunktiks on adekvaatse murdejoonte komplekti väljaselgitamine ning andmete ekstraheerimine ning lingvistiline analüüs: eesti murrete varieerumine on väga suur ning mitmetasandiline, eriti kui pöörata pilk (morfo)süntaktilisele varieerumisele. Näiteks eesti murrete vajadust või kohustust väljendavate *tarvis-* ja *vaja*-konstruktsioonide analüüsil selgus, et murretes varieerub konstruktsiooni iga osa (võtmata arvesse ka häälikulist varieerumist): [*mul/mulle/0*] [*on/tuleb/läheb/0*] [*vaja/tarvis*] [*uus arvuti/uut arvutit/kooli minna*]. Seejuures mõjutas valikut nii tähendus, murdeala kui ka konstruktsiooni mõni teine osa. (Lindström jt 2014; Lindström, Uiboaed 2017) Kobardialektoloogiline lähenemine eeldab suurt eeltööd varieerumise ulatuse ja variantide väljaselgitamisel; toodud näide *tarvis-/vaja*-konstruktsioonide varieerumisest võimaldaks juba omakorda kobardialektoloogilisi meetodeid rakendada.

5. Kokkuvõte

Tänapäeva dialektoloogia kasutab väga palju erinevaid meetodeid, mis nõuavad murdeuurijalt teadlikkust olemasolevatest metodoloogilistest võimalustest ja piirangutest. Meetodi valik sõltub ennekõike andmestikust ning eesmärgist; samuti sellest, kas ja kui palju võetakse lisaks keelelisele kaugusele arvesse geograafilist kaugust (geoinfot laiemalt) või muud tüüpi andmeid, mis võivad olla täiesti mittekeelelised. Kuna ruumiandmed on murdeandmete lahutamatu osa ning geograafiline mõõde tihtipeale oluline tegur keelenähtuste varieerumise seletamisel, on oluline ka andmete ja analüüsitulemuste kaartidel visualiseerimine.

Artiklis tutvustasime meetodeid, mis on otsapidi jõudnud ka eesti murdeuurimisse. Kui mõelda tulevikuperspektiividele, võib arvata, et huvi keelelise varieerumise ning (koha)murrete vastu keeleteaduses ei vaibu. Seda vaatamata sellele, et traditsiooniliste kohamurrete nivelleerumine toimub globaliseerumise tõttu praegu pea kõikjal maailmas ning ka keeleteaduses on sotsiaalsete murrete uurimine mõnevõrra väljapaistvamal positsioonil. Peab lootma, et ühes üldise huviga keele varieerumise vastu jääb püsima huvi varasema keelelise ja kultuurilise ajaloo vastu ning sellega seoses püsivad ja arenevad edasi ka artiklis mainitud meetodid ja suunad. Lisanduvad tõenäoliselt

meetodid ja tehnikad, mis võimaldavad paremini analüüsida urbaniseeruvas või virtuaalses keskkonnas aset leidvaid muutusi ning kujunevaid allkeeli, n-ö uusi murdeid.

Andmekeskne suund teaduses püsib ja kasvab ning vajadus nii vanade kui ka uute murrete andmete järele vaid suureneb. Üha enam vajame suhteliselt hõlpsalt kättesaadavaid digitaalsele kujule viidud andmeid, mida eri tüüpi analüüsides kasutada. Murdeerinevuste selgitamiseks ning murrete kujunemise modelleerimiseks vajame täienduseks kindlasti ka muud tüüpi infot: geograafilisi, kultuurilisi, rahvastikuandmeid jne, mis nõuavad interdistsiplinaarset koostööd. Nii andmete loomine, säilitamine, kogumine kui ka analüüs, sellega kaasnevad uued tehnilised oskused ja muutuvad mõtteviisid on väljakutseks nii dialektoloogiale, keeleteadusele kui ka (digi)humanitaariale laiemalt.

Artikli valmimist on toetanud Euroopa Liidu Regionaalarengu Fond (*Eesti-uuringute Tippkeskus*).

Võrguviited

ArchiMob = Archimob corpus of Swiss German, University of Zurich. https://www.spur.uzh.ch/en/departments/research/textgroup/ArchiMob.html (24. VIII 2018).

ASIt = Atlante Sintattico d'Italia, Università di Padova, Università di Venezia. http://asit.maldura.unipd.it (24. VIII 2018).

DynaSAND = Dynamic Syntactic Atlas of the Dutch dialects. Sjef Barbiers jt 2006. Amsterdam: Meertens Institute. http://www.meertens.knaw.nl/sand/ (24. VIII 2018).

EMK = http://www.murre.ut.ee/murdekorpus/ (24. VIII 2018).

EMSA = http://www.murre.ut.ee/arhiiv/ (24. VIII 2018).

EMSUKA = http://emsuka.eki.ee/ (24. VIII 2018).

FRED = Freiburg English Dialect Corpus. Albert-Ludwigs-Universität Freiburg. http://www2.anglistik.uni-freiburg.de/institut/lskortmann/FRED/ (24. VIII 2018).

GDC = Georgian Dialect Corpus. http://corpora.co/ (24. VIII 2018).

Lauseopin arkisto = Turun yliopisto, kieli- ja käännöstieteiden laitos, Kotimaisten kielten keskus, 1985. Lauseopin arkiston murrekorpuksen Helsinki-Korp-versio [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2016040702 (24. VIII 2018).

MKWEB = http://www.murre.ut.ee/mkweb/ (24. VIII 2018).

REDE = Regionalsprache.de. Akademie der Wissenschaften und der Literatur – Mainz. https://www.regionalsprache.de/en/Default.aspx (24. VIII 2018).

RuRaKe = http://rurake.keeleressursid.ee/index.php/dialect-maps/ (24. VIII 2018).

Kirjandus

A r i s t e, Paul 1956. Läänemere keelte kujunemine ja vanem arenemisjärk. – Eesti rahva etnilisest ajaloost. Toim Harri Moora. Tallinn: Eesti Riiklik Kirjastus, lk 5–23.

Baayen, R. Harald 2008. Analyzing Linguistic Data: A Practical Introduction to Statistics Using R. Cambridge: Cambridge University Press.
Breiman, Leo, Friedman, Jerome, Olshen, Richard A., Stone, Charles J. 1984. Classification and Regression Trees. Belmont, Calif.: Wadsworth.
Carrilho, Ernestina 2010. Tools for dialect syntax: The case of CORDIAL-SIN (An annotated corpus of Portuguese dialects). – Anuario del Seminario de Filología Vasca „Julio de Urquijo", nr 53, lk 57–70.
Cedergren, Henrietta J., Sankoff, David 1974. Variable rules: Performance as a statistical reflection of competence. – Language, kd 50, nr 2, lk 333–355.
Everitt, Brian S., Landau, Sabine, Leese, Morven, Stahl, Daniel 2011. Cluster Analysis. 5., parandatud ja täiendatud väljaanne. Chichester: Wiley-Blackwell.
Goebl, Hans 2018. Dialectometry. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Hoboken, NJ: Wiley-Blackwell, lk 123–142.
Grafmiller, Jason, Szmrecsanyi, Benedikt, Röthlisberger, Melanie, Heller, Benedikt (toim) 2017. Probabilistic Grammars: Syntactic Variation in a Comparative Perspective. – Special Collection. Glossa: A Journal of General Linguistics. https://www.glossa-journal.org/collections/special/probabilistic-grammars-syntactic-variation/ (24. VIII 2018).
Greenacre, Michael 2007. Correspondence Analysis in Practice. 2. tr. Boca Raton Fla.: CRC Press.
Gries, Stefan Th. 2009. Statistics for Linguistics with R. A Practical Introduction. Berlin: De Gruyter Mouton.
Grieve, Jack 2014. A comparison of statistical methods for the aggregation of regional linguistic variation. – Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech. Toim Benedikt Szmrecsanyi, Bernhard Wälchli. Berlin: Walter de Gruyter, lk 53–88.
Grieve, Jack 2018. Spatial statistics for dialectology. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Hoboken, NJ: Wiley-Blackwell, lk 415–433.
Heeringa, Wilber, Prokić, Jelena 2018. Computational Dialectology. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Hoboken, NJ: Wiley-Blackwell, lk 330–347.
Hennoste, Tiit, Muischnek, Kadri 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikooli Kirjastus, lk 183–317.
Hey, Tony, Tansley, Stewart, Tolle, Kristin M. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Kd 1. Redmond, WA: Microsoft Research.
Honkola, Terhi 2016. Macro- and Microevolution of Languages: Exploring Linguistic Divergence with Approaches from Evolutionary Biology. (Turun Yliopiston julkaisuja – Annales Universitatis Turkuensis. Ser. C AII.) Turku: Turun yliopisto.
Hoyningen-Huene, Paul 2018. The Human Sciences between Quantification and Hermeneutics. – Loeng Tartu Ülikoolis 6. II 2018.
Ikola, Osmo 1960. Perfektin ja pluskvamperfektin synnystä. – Virittäjä, kd 64, lk 364–368.

Janda, Laura A. (toim) 2013. Cognitive Linguistics – The Quantitative Turn: The Essential Reader. Walter de Gruyter.
Johannessen, Janne Bondi, Priestley, Joel, Hagen, Kristin, Åfarli, Tor Anders, Vangsnes, Øystein Alexander 2009. The Nordic Dialect Corpus – an advanced research tool. – Proceedings of the 17th Nordic Conference of Computational Linguistics NODALIDA 2009. (NEALT Proceedings Series 4.) Toim K. Jokinen, E. Bick. Odense: Northern European Association for Language Technology (NEALT), lk 73–80.
Juhkam, Evi 2012. Harju-Madise murrak. Toim Mari-Liis Kalvik, Helmi Neetar. Tallinn: Eesti Keele Sihtasutus.
Juhkam, Evi, Sepp, Aldi 2000. Läänemurde tekstid. (Eesti murded VIII.) Tallinn: Eesti Keele Instituut.
Kitchin, Rob 2014. The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences. Sage.
Klavan, Jane, Pilvik, Maarja-Liisa, Uiboaed, Kristel 2015. The use of multivariate statistical classification models for predicting constructional choice in spoken, non-standard varieties of Estonian. – SKY Journal of Linguistics, nr 28, lk 187–224.
Kretzschmar, William A. 2018. Linguistic Atlases. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Hoboken, NJ: Wiley-Blackwell, lk 57–72.
Krikmann, Arvo, Pajusalu, Karl 2000. Kus on keskmurde keskpunkt. – Inter dialectos nominaque. Pühendusteos Mari Mustale 11. novembril 2000. Toim Jüri Viikberg. Tallinn: Eesti Keele Sihtasutus, lk 131–172.
Laakso, Johanna 2001. The Finnic languages. – Circum-Baltic Languages, kd I: Past and Present. Toim Östen Dahl, Maria Koptjevskaja-Tamm. Amsterdam–Philadelphia: John Benjamins Publishing Company, lk 179–212.
Laanest, Arvo 1975. Sissejuhatus läänemeresoome keeltesse. Tallinn: Eesti NSV Teaduste Akadeemia Keele ja Kirjanduse Instituut.
Lebart, Ludovic, Salem, André, Berry, Lisette 1998. Exploring Textual Data. Dordrecht: Kluwer Academic Publishers.
Le Page, R. B. 1997. The evolution of a Sociolinguistic Theory of Language. – The Handbook of Sociolinguistics. Toim Florian Coulmas. Oxford: Blackwell, lk 15–32.
Levshina, Natalia 2015. How to do Linguistics with R: Data Exploration and Statistical Analysis. Amsterdam–Philadelphia: John Benjamins Publishing Company.
Lindström, Liina 2001. Eesti murrete korpuse iseloomustus argivestlusega võrrelduna. – Keele kannul. Pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. (Tartu Ülikooli eesti keele õppetooli toimetised 17.) Tartu: Tartu Ülikooli Kirjastus, lk 212–221.
Lindström, Liina 2015. Ülevaade eesti murrete korpusest seisuga 17.11.2015. https://www.keel.ut.ee/sites/default/files/www_ut/emk_teejuht2015.pdf (18. VIII 2018).
Lindström, Liina 2017. Partitive subjects in Estonian dialects. – ESUKA/JEFUL, kd 8, nr 2, lk 191–231.

Lindström, Liina, Kalmus, Mervi, Klaus, Anneliis, Bakhoff, Liisi, Pajusalu, Karl 2009. Ainsuse 1. isikule viitamine eesti murretes. – Emakeele Seltsi aastaraamat 54 (2008). Tallinn: Emakeele Selts, lk 159–185.
Lindström, Liina, Lonn, Varje, Mets, Mari, Pajusalu, Karl, Teras, Pire, Veismann, Ann, Velsker, Eva, Viikberg, Jüri 2001. Eesti murrete korpus ja kolme murde sagedasema sõnavara võrdlus. – Keele kannul: pühendusteos Mati Erelti 60. sünnipäevaks 12. märtsil 2001. (Tartu Ülikooli eesti keele õppetooli toimetised 17.) Tartu: Tartu Ülikooli Kirjastus, lk 186–211.
Lindström, Liina, Pilvik, Maarja-Liisa, Plado, Helen 2018. Nimetamiskonstruktsioonid eesti murretes: murdeerinevused või suuline süntaks? – Mäetagused, nr 70, lk 91–126.
Lindström, Liina, Pilvik, Maarja-Liisa, Ruutma, Mirjam, Uiboaed, Kristel 2015. Mineviku liitaegade kasutusest eesti murretes keelekontaktide valguses. – Aig õdagumeresoomõ keelin. Aeg läänemeresoome keeltes. (Võro Instituudi toimõndusõq 29.) Võro, lk 39–70.
Lindström, Liina, Pilvik, Maarja-Liisa, Ruutma, Mirjam, Uiboaed, Kristel 2018 (ilmumas). On the use of perfect and pluperfect in Estonian dialects: Frequency and language contacts. – Plurilingual Finnic. Change of Finnic languages in a multilinguistic environment. (Uralica Helsingiensia.) Toim Sofia Björklöf, Santra Jantunen. Helsinki: Finno-Ugrian Society.
Lindström, Liina, Uiboaed, Kristel 2017. Syntactic variation in 'need'-constructions in Estonian dialects. – Nordic Journal of Linguistics, kd 40, nr 3, lk 313–349.
Lindström, Liina, Uiboaed, Kristel, Vihman, Virve-Anneli 2014. Varieerumine *tarvis-/vaja*-konstruktsioonides keelekontaktide valguses. – Keel ja Kirjandus, nr 8–9, lk 609–630.
McEnery, Tony, Hardie, Andrew 2013. The history of corpus linguistics. – The Oxford Handbook of the History of Linguistics. Toim Keith Allan. Oxford: Oxford University Press, lk 727–746.
Mets, Mari 2010. Suhtlusvõrgustikud reaalajas: võru kõnekeele varieerumine kahes Võrumaa külas. (Dissertationes philologiae estonicae Universitatis Tartuensis 25.) Tartu: Tartu Ülikooli Kirjastus.
Metslang, Helle, Lindström, Liina 2017. Chapter 3. The essive in Estonian. – Uralic Essive and the Expression of Impermanent State. (Typological Studies in Language 119.) Toim Casper de Groot. Amsterdam: John Benjamins Publishing Company, lk 57–90.
Meyers, Lawrence S., Gamst, Glenn C., Guarino, Anthony J. 2006. Applied Multivariate Research: Design and Interpretation. Thousand Oaks: Sage Publications.
Murumets, Sirje 1982. Eesti keeleala murdelisest liigendusest „Väikese murdesõnastiku" põhjal I–II. – Keel ja Kirjandus, nr 1, lk 11–17; 1983, nr 11, lk 615–623.
Must, Mari 1987. Kirderannikumurre. Häälikuline ja grammatiline ülevaade. Tallinn: Valgus: Eesti NSV Teaduste Akadeemia, Keele ja Kirjanduse Instituut.
Neetar, Helmi 1964. Aluse ja öeldise ühildumist mõjutavatest teguritest eesti murretes. – Emakeele Seltsi aastaraamat X (1964). Tallinn: Eesti NSV Teaduste Akadeemia Emakeele Selts, lk 151–166.

Neetar, Helmi 1965a. Aluse ja öeldise ühildumise seaduspärasustest eesti murretes. – Keel ja Kirjandus, nr 1, lk 25–29.
Neetar, Helmi 1965b. Aluse (kollektiivsubstantiivi) ja öeldise ühildumisest eesti murretes. – Emakeele Seltsi aastaraamat 11 (1965). Tallinn: Eesti NSV Teaduste Akadeemia Emakeele Selts, lk 185–193.
Nerbonne, John, Kleiweg, Peter 2007. Toward a dialectological yardstick. – Journal of Quantitative Linguistics, kd 14, nr 2–3, lk 148–166.
Nerbonne, John, Kretzschmar Jr., William A. 2013. Dialectometry++. – Literary and Linguistic Computing, kd 28, nr 1, lk 2–12.
Nerbonne, John, Wieling, Martijn 2018. Statistics for Aggregate Variationist Analyses. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Hoboken, NJ: Wiley-Blackwell, lk 400–414.
Nigol, Salme 1994. Hargla murraku konsonantism. Toim Karl Pajusalu. Tallinn: Eesti TA Eesti Keele Instituut.
Nurkse, Rein 1937. Adjektiiv-atribuudi kongruentsist eesti keeles. (Akadeemilise Emakeele Seltsi toimetused 30.) Tartu: Akadeemilise Emakeele Seltsi Kirjastus.
Pajusalu, Karl, Velsker, Eva, Org, Ervin 1999. On recent changes in South Estonian: Dynamics in the formation of the inessive. – International journal of the Sociology of Language, kd 139, nr 1, lk 87–104.
Paolillo, John C. 2002. Analyzing Linguistic Variation. Statistical Models and Methods. Stanford: CSLI Publications.
Pilvik, Maarja-Liisa 2017. Deverbal *-mine* action nominals in the Estonian dialect corpus. – ESUKA/JEFUL, kd 8, nr 2, lk 295–326.
Plado, Helen 2015. *des-* ja *mata*-konverbi kasutusest eesti murretes. – Emakeele Seltsi aastaraamat 60 (2014). Tallinn: Teaduste Akadeemia Kirjastus, lk 195–218.
Pook, Hanna 2018. Pronoomeni *kes* kasutusest eesti murretes. Magistritöö. Tartu: Tartu Ülikool. http://hdl.handle.net/10062/60630
Rankin, Robert L. 2003. The comparative method. – The Handbook of Historical Linguistics. Toim Brian D. Joseph, Richard D. Janda. Oxford: Blackwell, lk 183–212.
Ruutma, Mirjam 2016. Kaassõnad eesti murretes. Magistritöö. Tartu: Tartu Ülikool. http://hdl.handle.net/10062/51736
Ruutma, Mirjam, Kyröläinen, Aki-Juhani, Pilvik, Maarja-Liisa, Uiboaed, Kristel 2016. Ambipositsioonide morfosüntaktilise varieerumise kirjeldusi kvantitatiivsete profiilide abil. – Keel ja Kirjandus, nr 2, lk 92–113.
Rätsep, Huno 2003. Tartu ülikooli eesti keele arhiivi saamisloost ja saatusest. – 200 aastat eesti keele ülikooliõpet: 1803 eesti ja soome keele lektoraat Tartu ülikoolis. (Tartu Ülikooli eesti keele õppetooli toimetised 25.) Toim Mati Erelt, Valve-Liivi Kingisepp. Tartu: Tartu Ülikooli Kirjastus, lk 153–170.
Saareste, Andrus 1938. Eesti murdeatlas. I vihk. Tartu: Eesti Kirjanduse Selts.
Saareste, Andrus 1941. Eesti murdeatlas. II vihk. Tartu: Eesti Kirjanduse Selts.
Saareste, Andrus 1955. Petit atlas des parlers estoniens. Väike eesti murdeatlas. Uppsala: Almqvist & Wiksells.
Séguy, Jean 1973. La dialectométrie dans l'Atlas linguistique de la Gascogne. – Revue de linguistique romane, kd 37, nr 145–146, lk 1–24.

Serebrennikov, B. A. 1959. Pluskvamperfekti ja perfekti päritolu probleemist läänemeresoome keeltes. – Emakeele Seltsi aastaraamat IV (1958). Tallinn: Eesti Riiklik Kirjastus, lk 249–255.
Szmrecsanyi, Benedikt 2013. Grammatical Variation in British English Dialects: A Study in Corpus-Based Dialectometry. Cambridge: Cambridge University Press.
Szmrecsanyi, Benedikt 2014. Forests, trees, corpora, and dialect grammars. – Aggregating Dialectology, Typology, and Register Analysis: Linguistic Variation in Text and Speech. Toim B. Szmrecsanyi, Bernhard Wälchli. Berlin–Boston: Walter de Gruyter, lk 89–212.
Szmrecsanyi, Benedikt, Anderwald, Lieselotte 2018. Corpus-Based Approaches to Dialect Study. – The Handbook of Dialectology. Toim Charles Boberg, John Nerbonne, Dominic Watt. Hoboken, NJ: Wiley-Blackwell, lk 300–313.
Szmrecsanyi, Benedikt, Grafmiller, Jason, Heller, Benedikt, Röthlisberger, Melanie 2016. Around the world in three alternations: Modeling syntactic variation in varieties of English. – English World-Wide, kd 37, nr 2, lk 109–137.
Szmrecsanyi, Benedikt, Grafmiller, Jason, Bresnan, Joan, Rosenbach, Anette, Tagliamonte, Sali, Todd, Simon 2017. Spoken syntax in a comparative perspective: The dative and genitive alternation in varieties of English. – Glossa: A Journal of General Linguistics, kd 2, nr 1, artikkel 86.
Strobl, Carolin, Malley, James, Tutz, Gerhard 2009. An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. – Psychological Methods, kd 14, nr 4, lk 323–348.
Syrjänen, Kaj, Honkola, Terhi, Lehtinen, Jyri, Leino, Antti, Vesakoski, Outi 2016. Applying population genetic approaches within languages: Finnish dialects as linguistic populations. – Language Dynamics and Change, kd 6, nr 2, lk 235–283.
Tagliamonte, Sali A. 2013. Analysing and interpreting variation in the sociolinguistic tradition. – Research Methods in Language Variation and Change. Toim Manfred Krug, Julia Schlüter. Cambridge: University Press, lk 382–401.
Tagliamonte, Sali A., Baayen, R. Harald 2012. Models, forests, and trees of York English: *Was/were* variation as a case study for statistical practice. – Language Variation & Change, kd 24, nr 2, lk 135–178.
Tooding, Liina-Mai 2015. Andmete analüüs ja tõlgendamine sotsiaalteadustes. Teine, täiendatud väljaanne. Tartu: Tartu Ülikooli Kirjastus.
Uiboaed, Kristel 2013. Verbiühendid eesti murretes. (Dissertationes philologiae estonicae Universitatis Tartuensis 34). Tartu: Tartu Ülikooli Kirjastus.
Uiboaed, Kristel, Hasselblatt, Cornelius, Lindström, Liina, Muischnek, Kadri, Nerbonne, John 2013. Variation of verbal constructions in Estonian dialects. – Literary & Linguistic Computing, kd 28, nr 1, lk 42–62.
Univere, Aili 1988. Idamurre. – Emakeele Seltsi aastaraamat 32 (1986). Tallinn: Eesti Raamat, lk 59–93.
Velsker, Eva 2000. Inessiivi lõpu varieerumine Vastseliina murrakus. Magistritöö Tartu Ülikooli eesti keele osakonnas.

Walker, James A. 2013. Variation analysis. – Research Methods in Linguistics. Toim Robert J. Podesva, Devyani Sharma. Cambridge: University Press, lk 440–459.
Wolk, Christoph, Szmrecsanyi, Benedikt 2016. Top-down and bottom-up advances in corpus-based dialectometry. – The Future of Dialects. Selected papers from Methods in Dialectology XV. Toim Marie-Hélène Côté, Remco Knooihuizen, John Nerbonne. Berlin: Language Science Press, lk 225–244.

Liina Lindström (*snd 1973*)*, PhD, Tartu Ülikool, eesti ja üldkeeleteaduse instituut, eesti keele dotsent, liina.lindstrom@ut.ee*

Maarja-Liisa Pilvik (*snd 1989*)*, doktorant, Tartu Ülikool, eesti ja üldkeeleteaduse instituut, rakendusliku dialektoloogia nooremteadur, maarja-liisa.pilvik@ut.ee*

Corpus-based quantitative dialectology

Keywords: Estonian dialects, research methods, dialect corpus, variation studies

The article gives an overview of the directions and trends in dialectology with an emphasis on Estonian dialectology. We compare different methods and approaches for studying local language varieties: traditional dialectology based on the historical-comparative method, atlas-based dialectology, variation studies which stem from variationist sociolinguistics, and corpus-based approaches, which have been gaining momentum in recent years thanks to the compilation and development of the Estonian Dialect Corpus. In the article, we give an overview of the type of data and methods these approaches use. While traditional dialectology collected abundant qualitative data, which were based on texts and questionnaires, in order to compile dialect descriptions, dictionaries and atlases, newer, corpus-based methods use frequency data obtained from the corpus for comparing the dialects, modeling the variation, and examining aggregate distributions of linguistic phenomena in the corpus. The latter means comparing dialects and their linguistic distances on the basis of analysing the distributions of many linguistic features. The methodology used in corpus-based quantitative dialectology is rich and constantly developing, enabling the researcher to account for more and more aspects underlying linguistic variation.

Liina Lindström (*b. 1973*)*, PhD, University of Tartu, Institute of Estonian and General Linguistics, Associate Professor of Estonian Language, liina.lindstrom@ut.ee*

Maarja-Liisa Pilvik (*b. 1989*)*, PhD Student, University of Tartu, Faculty of Arts and Humanities, Institute of Estonian and General Linguistics, Junior Research Fellow, maarja-liisa.pilvik@ut.ee*