Regilaulude teema-analüüs: võimalusi ja väljakutseid[1]

Mari Sarv

Teesid: Artikkel uurib regilaulu teema-analüüsi võimalusi teemade modelleerimise meetodi abil. Meetodi kasutamisel on probleemiks regilaulu keele piirkondlik varieeruvus. Laulutekstide esmane analüüs näitas, et sisukamaid tulemusi annab teema-analüüs ühtlasema keelega kogumite puhul. Lähemaks vaatluseks valitud Hiiumaa, Saaremaa ja Muhu laulude teema-analüüsil tuvastati 20 teemat, mis annavad kiire ülevaate vaadeldavate laulude temaatilisest struktuurist. Uurimus näitas, et tuvastatud teemad jaotuvad vaadeldud piirkonnas võrdlemisi ühtlaselt. Kuid arvutuslikud teemarühmad ei kattu üheselt regilaulu varasema liigitusega, arvestamata laulude žanrilisi erinevusi ning tuues esiplaanile vaadeldavas laulukogumis sagedamini esinevad laulutüübid.

DOI: 10.7592/methis.v21i26.16914

Võtmesõnad: regilaul, teemade modelleerimine, korpuspõhine analüüs, folkloor, digihumanitaaria

Artikli eesmärgiks on uurida eesti regilaulutekstide arvutusliku teema-analüüsi võimalusi teemade modelleerimise meetodi abil. Meetod võimaldab eristada valitud tekstikogumis abstraktsed teemad, mis leitakse puht-statistiliselt selle alusel, kui sageli sõnad samas tekstis üheskoos esinevad. Regilaulu puhul võiks sellisest analüüsist abi olla suurte laulukoguste temaatilise struktuuri tuvastamisel ja piirkonnatraditsioonide võrdlemisel, aga ka laulude klassifitseerimisel ja seniste liigituspõhimõtete analüüsimisel. Regilaulutekstide arvutusliku analüüsi juures on väljakutseks laulude suur keeleline varieeruvus ja korpuse ebaühtlus nii kogumistiheduse kui eri laulutüüpide esindatuse osas. Need asjaolud tekitavad küsimuse, kas sellise tekstikogumi teema-analüüsiga on üldse võimalik laulude sisutasandini jõuda ning kui mõttekas selline analüüs oleks. Siinne artikkel vahendab esimeste katsetuste tulemusi selles vallas.

Regilaulude korpuspõhise uurimise võimalused ja väljakutsed

Regilaul, eesti kultuuri üks eripärasemaid nähtusi, moodustab osa läänemeresoomlaste ühisest laulutraditsioonist. Regilaule ja teavet selle sisu ja kasutuse kohta

1 Uurimuse valmimist on toetanud Haridus- ja Teadusministeerium (projekt EKKD65 „Kuidas allikatest saab kultuur: eesti aines Eesti Kirjandusmuuseumi kogudes ja andmebaasides"), Soome Akadeemia (projekt nr 333138 „Vormellik intertekstuaalsus, teemavõrgustikud ja poeetiline varieerumine läänemeresoome suulise luule piirkonnatraditsioonides"), Eesti Kirjandusmuuseum (baasrahastusprojekt EKM 8-2/20/2) ja Euroopa Liit Euroopa Regionaalarengu Fondi kaudu (Eesti-uuringute Tippkeskus).

on kogutud, kirja pandud või salvestatud eri aegadel ning põhiline osa sellest teabest on koondatud Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiivi, mille ülesandeks on tagada selle teabe säilimine, kättesaadavus ning uurimise, tänapäevase mõtestamise, teavitus- ja haridustöö kaudu selle hoidmine meie kultuurikäibes ka edaspidi. Regilaulutekstide avalikkusele kättesaadavaks tegemiseks ja nende paremaks uurimiseks on alates 2003. aastast koostatud rahvaluulearhiivis eesti regilaulude andmebaasi (Oras, Saarlo ja Sarv 2003–2020). Praeguseks sisaldab andmebaas ligikaudu 2/3 kõigist Eestis kogutud regilaulutekstidest, s.o umbes 100 000 teksti koos põhilise teabega kogumise aja ja koha, koguja ja esitaja kohta (juhul kui vastav info on dokumenteeritud).

Suure osa regilaulutekstide olemasolu digitaalkujul avab meile uusi võimalusi regilaulutraditsiooni uurimiseks. Digieelsel ajastul oli väga raske, kui mitte võimatu süstemaatiliselt läbi töötada suuri tekstikogumeid, et saada ülevaatlikke ja samas tõepäraseid teadmisi eesti regilaulu kohta. Nii näiteks on laialdaselt käibele läinud (ja isegi kooliõpetusse jõudnud) ekslik väide, et eesti regilaulus kehtivad kalevalamõõdu kvantiteedireeglid (vt nt Aavik 1914; Laugaste 1977, 144–145; Tedre 1998, 550). Värsianalüüsi ajamahukus ei võimaldanud seda väidet enne digiajastu saabumist kontrollida, kuid nüüdseks teame, et läänemeresoome regilaulude hulgas on eesti regilaulud nimelt sellised, kus kvantiteedireeglid pigem ei kehti (vt lähemalt Sarv 2015). Praegu, mil andmebaasi on lisatud juba kõige olulisemad ja mahukamad regilaulukogud, on mõttekas kasutada ära võimalus saada selle põhjal teadmisi eesti regilaulutraditsiooni kohta, mis võtaks arvesse kogu regilaulu andmestikku, ning kõrvutada neid varasemate uurimistulemustega (nt Sarv 2019). Nii näiteks on arvutianalüüsiga võimalik selgitada, kuidas, milliste tunnuste põhjal on kujunenud piirkondlikud traditsioonid ja kus kulgevad regilaulupiirkondade eraldusjooned, uurida regilaulu arhailise keele erijoonte levikut jpm.

Regilaulude analüüsil peame arvestama, et varieeruvus, mis on üks folkloori põhitunnuseid, avaldub ja on levinud folkloori eri aspektide puhul erinevalt – sisu, vormi, esitust ja kasutust iseloomustavad jooned järgivad erisuguseid levikumustreid. Meie rahvaluulekogude tekstide arvutianalüüsil tuleb silmas pidada, et folkloorse varieerumise aluskihiks on alati keeleline varieerumine – folklooritekstid on enamjaolt kirja pandud murdekeeles, kus varieerumine hõlmab keele eri aspekte (sõnavara, morfoloogia, süntaks jm). Lisaks sellele ei ole arhiivi kogutud materjal jaotunud võrdselt ei teemade ega ka piirkondade lõikes. Nagu Arvo Krikmann on näidanud, kehtib Georg Kingsley Zipfi poolt keelekasutuse kohta formuleeritud seadus, mis lihtsustatult öeldes näitab, et keeles kasutatakse väga väheseid sõnu väga sageli

ja väga paljusid sõnu väga harva, ka rahvaluulekogude tüpoloogilise jaotuse[2] kohta (Krikmann 1997). Samuti erineb regilaulude kogumistihedus piirkonniti märkimisväärselt, mõistetavatel põhjustel on rohkem regilaule jäädvustatud kogumisajal elujõulisema ja vanapärasema traditsiooniga kihelkondadest ja vähem piirkondadest, kus regilaulutraditsioon on varem kaduma hakanud või olnud vähem tuntud (näiteks eesti-rootsi sega-asustusega alad). Kogu korpuse statistilisel vaatlusel tuleb meeles pidada, et andmestikus kipuvad sagedased tüübid ja suurema tekstiarvuga esindatud kihelkonnad domineerima. Neid näitajaid on võimalik ühtlustada, kuid ka see pole alati sobiv lahendus, kuna annaks teistpidi moonutatud tulemuse, milles tõuseksid ebaproportsionaalselt esile haruldasemate tüüpide või hõredama regilaulutraditsiooniga alade tekstid.

Digihumanitaarias on loodud ja kasutusele võetud mitmeid võimalusi tekstide statistiliseks uurimiseks. Tihti eeldavad keelestatistikal põhinevad meetodid, et sõnad on lemmatiseeritud, s.t sõna eri esinemiskujud ja vormid on kokku viidud. Eriti oluline on see tekstide sisu kohta käivates uuringutes ning keeltes, kus ühel sõnal võib olla palju eri vorme, nagu see on läänemeresoome keeltes. Väga paljude keelte jaoks on loodud automaatsed keeleanalüüsi vahendid, mida saab kasutada tekstisõnade grammatiliseks analüüsiks ja lemmatiseerimiseks. Selliste vahendite loomine ongi tavaliselt mõistlik ja vajalik standardiseeritud kirjakeelega keelte puhul, kus kasutatavad sõnavormid väga palju ei varieeru. Varieeruva murdekeele jaoks, ja veelgi enam, rohkete eripäraste arhaismidega regilaulukeele jaoks keegi sellist analüsaatorit seni veel loonud ei ole, kuna selles keeles tekstide maht piirdubki peamiselt rahvaluulearhiivi kogudega ning tekstide massilist lisandumist tulevikus oodata ei ole. Regilaulutekstide analüüsimine sõnahaaval, nagu seda on tehtud näiteks Tartu Ülikooli eesti murrete korpuse loomisel (vt Lindström 2015, 10), annaks küll täpse ja usaldusväärse tulemuse, kuid oleks äärmiselt ajamahukas tegevus. Keeleanalüüsi vahendid ja meetodid arenevad kiiresti ja küllap oleks ka juba praegu suur osa regilaulu sõnedest võimalik ära analüüsida, kuid paratamatult oleks tulemus eri murdealadel ebaühtlane ja valiidsema tulemuse saamine nõuaks regilaulu- ja murdekeeletundja olulist ajalist panust sellesse tegevusse. Kuna lemmatiseeritud ja morfoloogiliselt analüüsitud regilaulutekstid oleks teadlastele siiski äärmiselt huvitav uurimismaterjal ja see teave oleks abiks ka regilauluandmebaasi otsinguvõimaluste parandamisel, siis loodame tulevikus sobiva lahenduseni kindlasti jõuda. Siiski on juba praegu suur soov saada uusi teadmisi meie regilaulutraditsiooni kohta, kasutades

2 Tüüp on folkloristikas kasutatav mõiste, mis hõlmab kindla sisu ja vormiga folkloorinähtuse ideed ja selle võimalikke esitusi (variante), vt nt Tedre 1974. Näiteks moodustavad ühe jututüübi muinasjutu „Punamütsike" idee ja selle kõik võimalikud esitused.

ära mahuka digikorpuse olemasolu ja võimalusi saada sellest arvuti abil uut laadi teavet. Regilauluandmebaasi tüpoloogiline korraldamine ei ole praegu veel nii kaugele jõudnud, et saaksime teha üldistavat statistikat eesti regilauluvarast ja öelda, kui suure osa üks või teine temaatiline või funktsionaalne laulurühm mingi piirkonna lauludest moodustab ning millised on temaatilise jaotuse piirkondlikud iseärasused, ning siin võiks tulla appi sõnastatistikal põhinev arvutuslik lähenemine.

Teemade modelleerimine: meetodi tutvustus

Teemade modelleerimine on meetod, millega on võimalik tuvastada uuritava tekstikogumi eri tekstides (või tekstide osades) korduvad sõnakogumid ehk abstraktsed, statistiliselt leitud „teemad". Meetod võimaldab saada kiiresti ülevaate vaadeldava tekstikogumi sisulis-temaatilisest koostisest ning on tänapäeval laialt kasutusel, et arvuti jõul analüüsida suuri ja väiksemaid andmehulki, mida eri elualadel meie käsutusse järjest rohkem ja rohkem lisandub. Arvutusliku teema-analüüsiga hakati tegelema 1990. aastatel ning aja jooksul on välja töötatud mitmesuguseid teema-analüüsi võimalusi, levinuimaks teemade modelleerimise algoritmiks tänapäeval on LDA (latentne Dirichlet' jaotamine – Blei jt 2003). Piiratud arvu tekstide põhjal treenitud mudelit on hiljem võimalik kasutada suurema hulga (või tulevikus lisanduvate) tekstide kategoriseerimiseks (teema-analüüsi meetodeid on täpsemini ja lugejasõbralikult lahti seletatud nt Topic Analysis 2020 ja Mäkelä 2018).

Arvuti poolt tuvastatud abstraktsed teemad on, nagu öeldud, sõnakogumid, mis korduvad üheskoos eri tekstides. Kui ajalehetekstides esinevad koos näiteks sõnad *võrkpall*, *kohtunik*, *söötis*, võime arvata, et tegemist on sporditeemaliste artiklitega, kui aga *peaminister*, *eelarve*, *majandus* ja *meetmed*, siis ilmselt poliitikateemaliste artiklitega. Kuna teema-analüüsiga soovitakse tavaliselt liigitada tekste nende sisu põhjal, jäetakse analüüsist kõrvale niinimetatud stoppsõnad ehk grammatilisema funktsiooniga sõnad, mis teksti sisu kohta ütlevad vähe (vt Uiboaed 2018). Algoritmi rakendamine tekstikogumile genereerib iga kord unikaalse ja teistest rakenduskordadest veidi erineva teemajaotuse. Arvuti teemade sisu ei mõista, analüüsitulemuste interpreteerimiseks on enamasti vajalik nende sildistamine ehk sõnakogumitele nimetuse andmine. Ka seda tööetappi on võimalik juba varem kogunenud digitaalse teabe alusel automatiseerida, kuid tihti tehakse seda siiski teema võtmesõnade põhjal inimjõul.

Enne oma uurimisküsimuste juurde asumist püüdsin selgitada, kuivõrd on arvutuslikku teema-analüüsi eesti ainesel, näiteks humanitaarteadustes, juba kasutatud. Eesti Teaduse Infosüsteem (ETIS) koondab lisaks selle muudele funktsioonidele ka teabe Eesti teadlaste poolt või nende osalusel valminud publikatsioonide kohta. ETISe andmetel näib teemade modelleerimine Eesti teaduses üsna haruldane meetod ole-

vat: märksõnade *topic modeling* ja *topic modelling* otsing andis kokku kolm kirjet, neist kahed konverentsiteesid; eestikeelne termin *teemade modelleerimine* on ETISe jaoks seni veel tundmatu.[3] Võrdluseks, rahvusvahelises teaduspublikatsioonide andmebaasis Google Scholar on samade märksõnadega seotud kokku 60000 kirjet, lisaks annab 20000 kirjet märksõna *topic models*, mis ETISest üldse puudub. Google Scholari otsing „teemade modelleerimine" andis tulemuseks kaheksa vastet, selgub et meetodit on eesti keeles kirjutavate teadlaste poolt kasutatud või selle kasutatavust hinnatud isiksusetestide, maanteeameti e-kirjade, pahaloomuliste kasvajarakkude DNA mutatsioonide ja sarnaseid tooteid arendavate ettevõtete analüüsimisel. Internetiotsinguga tuleb välja veel mitmeid eri mahu ja tasemega katsetusi ja uuringuid (näiteks on tuvastatud peamised teemad A. H. Tammsaare teoses „Tõde ja õigus" ja analüüsitud eri teemade käsitlemist erakondade valimisprogrammides, vt Eilat 2019 ja Mölder 2019). Üheks värskemaks teema-analüüsi rakenduseks on Maria Ivanova ettekanne vene kirjanduslike ballaadide teemade modelleerimisest ajaloo, kirjanduse ja kultuuriteaduste tudengikonverentsil (Ivanova 2020).

Regilaulutekstide teema-analüüs: eesmärk, uurimismaterjal, meetod

Minu uurimuse eesmärgiks on selgitada, kas teemade modelleerimise meetodiga on üldse võimalik saada sisukaid ja mõtestatavaid tulemusi regilaulude varieeruva ja lemmatiseerimata tekstikogumi kohta, täpsemalt, (1) kas sellest meetodist võiks olla abi regilaulukorpuse temaatilise struktuuri ja selle piirkondlike eripärade kirjeldamisel ning (2) kuidas suhestuks arvutuslikult saadud temaatiline jaotus varasema, regilaulude teadusliku liigitamise ja tüpologiseerimise käigus saadud jaotusega. Kasutan seda meetodit oma töös esimest korda, seega on ka minu enda jaoks tegemist meetodialase avastusretkega, mille käiku ja tulemusi siin vahendan.

Uurimismaterjalina kasutasin tekste pidevalt täienevast „Eesti regilaulude andmebaasist" (Oras, Saarlo ja Sarv 2002–2020) 2018. aasta novembri seisuga. Regilaulude andmebaasi puhul ei ole tegemist lõpetatud andmebaasiga, vaid andmete lisamine ja korrastamine toimub pidevalt. Andmebaasi lisatakse originaaliga võrreldud ja ortograafiliselt ühtlustatud tekstid, kuid laulude liigitus pärineb suuresti andmebaasi loomise aluseks olnud masinakirjakoopiatelt, mida on praeguseks üle vaadata jõutud ainult osaliselt. Seepärast ei ole andmed laulude liigilise ja tüübilise kuuluvuse kohta täiesti täpsed ja võivad sisaldada eksitavat teavet. Valisin oma vaatluse jaoks andmebaasist tekstid žanrimääratlusega „ainult regilaul" ootusega, et need

3 Eesti Teaduse Infosüsteemi administraatori Priit Tuvikese andmetel hõlmab ETISe publikatsioonide otsingu vormi lahter „Otsisõna" päringuid nii publikatsioonide pealkirjadest kui publikatsioonidele lisatud märksõnadest – e-kiri 12.11.2020.

tekstid sisaldaks võimalikult vähe muude rahvaluuležanride või laulustiilide elemente, mis võiks tulemust hägustada.

Tekstikogumist teemade tuvastamiseks kasutasin rakendust MALLET (McCallum 2002) ning levinuimat LDA algoritmi. Rakendusele tuleb anda tekstid ja soovitud teemade arv, täiendavalt võib lisada eri parameetreid: kas soovitakse koostada teemad nii, et nende osakaal kogumis oleks võrdne, või lähtutakse teemade loomulikust jaotusest (valisin viimase); kas eemaldatakse stoppsõnad või mitte jm (juhendit rakenduse MALLET kasutamiseks vt Graham jt 2012). Masina poolt arvutuslikult leitavate teemade arvu (ehk selle, mitmesse eri kogumisse arvuti sõnad jagab) määrab uurija; sageli selgub alles katsetuste käigus, milline arv teemasid konkreetse, uurimiseks valitud tekstikogumi puhul annab kõige sisukamaid tulemusi. Analüüsi väljundiks on (1) põhiinfo teemade kohta, s.t iga teema osakaal tekstikogumis ning olulisemate sõnade (võtmesõnade) loend; (2) iga teksti protsentuaalne jaotumine teemade vahel – üks tekst võib sisaldada ja enamasti sisaldabki sõnu rohkem kui ühest teemast; ning (3) teksti kõigi sõnede loend koos iga sõne asukohaandmete ja talle omistatud teema numbriga.

Sobiva mahuga ainekogumi valik

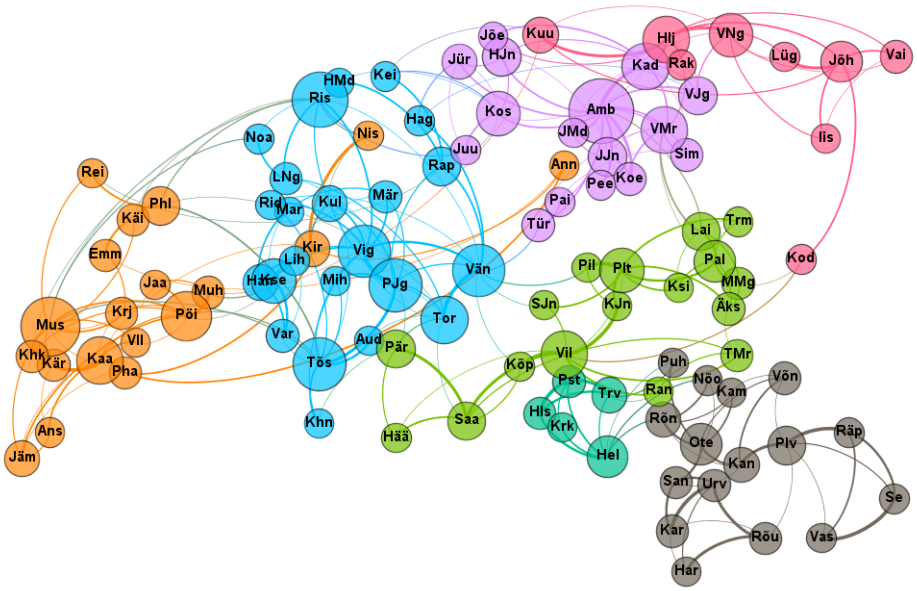
Otsustasin rakendada teema-analüüsi kõigepealt väikesele arvule failidele, et andmestik ja kuvatavad tulemused oleks minu jaoks hoomatavates mõõtmetes, ühtlasi aga saaksin hinnata meetodi toimivust ja ainese eripära tõttu esile kerkivaid probleeme. Koondasin kõik uuritavad regilaulutekstid maakondlikesse kogumitesse[4] (14 faili) ja lasin rakendusel tuvastada nendest 20 teemat. Tulemuste vaatlemisel selgus, et teemade määramisel olulisel kohal (ehk võtmesõnade hulgas) on palju grammatilise funktsiooniga sõnu ning sisu alusel teemade tuvastamiseks oleks kasulik siiski stoppsõnad eemaldada. Et grammatilise funktsiooniga sõnad esinevad keelekasutuses üleüldiselt sagedamini kui sisusõnad, oli võrdlemisi lihtne sõnavormide sagedusloendi alusel koostada sagedasemate stoppsõnade loend, ning edaspidi jätsin need teema-analüüsist kõrvale. Selgus, et laulukogumi teemadeks jaotamisel omandavad sellisel juhul olulise positsiooni lauludes esinevad refräänsõnad. Ka selgus teemade võtmesõnade vaatlusel, et teemade arvutamist on oluliselt mõjutanud keelemurre, kuus teemat kahekümnest olid selgelt lõunaeestikeelsed.

4 Folklooriainese arhiveerimisel ja liigitamisel lähtutakse üldiselt 19. sajandi lõpu administratiivsest jaotusest (kihelkonnad, maakonnad). Maakonnapiiridest sõltumatult paigutasin eesti keeleala peamisest jaotusest lähtuvalt eraldi kogumisse lõunaeestikeelse Mulgi ala laulud (maakondlikult jagatud Viljandi- ja Pärnumaa vahel) ning eraldi kogumitesse ka Lõuna- ja Põhja-Tartumaa laulud.

Järgmise katsetuse tegin keeleliselt mõnevõrra ühtlasemate põhjaeestikeelsete lauludega (8 maakonnakogumit), kus stoppsõnade loendisse olid lisatud (ehk teemajaotusest eemaldatud) ka tavalisemad refräänsõnad. Kaheksa põhjaeesti keeleala maakondliku tekstikogumi teema-analüüs genereeris teemad siiski nii, et eri teemades domineerisid eri murdealade regilauluidioomidesse kuuluvad sõnavaralised, häälikulised ja grammatilised vormid. Nii näiteks sisaldasid 20 teemast üheteist-kümne võtmesõnad mõiste „ema" eri variante ja vorme: *ema* (3 teemas), *emä* (2), *emakene* (1), *eit* (3), eide (3), *eite* (2), *eidekene* (4), *memme* (2), *memmekene* (2). Sellest esmasest vaatlusest oli ilmne, et regilaulude teema-analüüsi tulemusi kujundavad lisaks sisutunnustele laulutekstide murdelised erinevused – nagu oli ootuspärane arvata, takistas arvutuslike vahenditega regilaulu sisu juurde pääsemist laulude keeleline varieeruvus. Seega leidsin, et teema-analüüsi oleks mõttekam rakendada keeleliselt homogeensematele tekstidele.

Järgmiseks ülesandeks regilaulude sisuni jõudmisel oligi tuvastada regilaulukorpuses keeleliselt ühtlase(ma)d alad, mille piires teema-analüüs sisukaid tulemusi annaks. Teadupärast ei varieeru regilaulukeel mitte üksnes murdeliselt (Peegel 2006, 198–199), vaid regilaulukeele morfoloogiat kujundavad oluliselt ka värsimõõdu eripärad, kusjuures murde- ja värsimõõdualad omavahel üheselt kokku ei lange (Sarv 2019). Otsustasin ka selle ülesande lahendamisel kasutada arvuti abi, nimelt stilomeetrilist analüüsi, mis põhineb (praeguseks juba korduvalt kontrollitud) väitel, et igal inimesel on oma personaalne stiil, mis ilmneb statistiliselt tema poolt kõige sagedamini kasutatavate sõnade (vm üksuste) sagedusmustris. Nende sagedusmustrite võrdlemisel saab teha kindlaks eri tekstide omavahelise keelekasutusliku (ehk stilistilise) läheduse või kauguse. Stilomeetria esmaseks rakendusalaks on autorsuse tuvastamine, kuid selle meetodi abil on võimalik uurida ka keelekasutuse muutumist ja varieerumist laiemalt.

Teostasin regilaulukorpuse kihelkonnakogumite stilomeetrilise analüüsi rakendusega Stylo (Eder jt 2013) ning selle väljundina saadud võrgustikuandmetele rakendasin võrgustikuanalüüsi programmis Gephi (Bastian jt 2009) kasutatavat modulaarsusanalüüsi (Blondel jt 2008, vt ka Barabási 2016, peatükk 9), mis jagas Eesti kihelkonnad regilaulutekstide sagedasemate sõnade kasutuse mustrite sarnasuse alusel sarnase keelekasutusega kogumiteks (joonis 1). Edasiseks teema-analüüsiks valisin ühe sidusa piirkonna, Lääne-Eesti saarte, täpsemalt Hiiumaa, Saaremaa ja Muhu regilaulud (ma ei kaasanud analüüsi kogukonna põhiosast geograafiliselt eemal paiknevate üksikute kihelkondade Anna, Kirbla ja Nissi laule).



Joonis 1. Eesti regilaulualad regilaulutekstide sagedasemate sõnavormide kasutusmustrite sarnasuste põhjal (kihelkonnad on paigutatud kihelkondade geograafiliste koordinaatide alusel, joone jämedus kajastab sõlmede omavaheliste seoste tihedust, võrgustikus omavahel tihedamalt seotud sõlmekogumid on värvitud sama värvi).[5]

Hiiumaa, Saaremaa ja Muhu laulude teema-analüüsi tulemused

Teema-analüüsiks kasutasin regilauluandmebaasis olevaid Hiiumaalt, Saaremaalt ja Muhust kogutud laule (žanrimääratlusega „ainult regilaul"), kokku 3673 teksti.[6] Seekord valisin analüüsitavaks laulu tasandi (erinevalt esialgsel vaatlusel kasutatud maakonnakogumitest) nii, et iga laul moodustas omaette analüüsteksti. Lasin programmil tuvastada tekstidest, millest olid eemaldatud sagedasemad stoppsõnad, 20 teemat, mille osakaalud ei olnud ette antud, vaid järgisid teemade loomulikku jaotust materjalis. Esitan siinkohal analüüsi tulemusel saadud teemade tõenäosusnäitajad, võtmesõnade loendi ja sildi, mille panin võtmesõnade alusel ning mis viitab võtmesõnadega seostuvale regilaulude rühmale, mitmel juhul ka konk-

5 Selle kaardi üksikasjalikum analüüs ei kuulu käesoleva artikli eesmärkide hulka. Olgu ainult mainitud, et (1) siin toodud geograafilise paigutusega võrgustikujoonisel ei tule ilmsiks, et võrgustikus eristuvad omavahel selgelt lõunaeesti ja põhjaeesti murded, ja (2) analüüsi olid kaasatud refräänsõnad, mis oma sagedase esinemuse tõttu lõunapoolse Eesti regilauludes mõjutasid ka sarnasusnäitajaid ja võrgustiku kogukondadeks jaotumist, kuid peegeldavad pigem kultuurilist kui keelelist omapära. See puudutab eelkõige lõunaeesti keeleala ja sellega piirnevat põhjaeesti ala, aga ei tohiks oluliselt mõjutada siinses artiklis edaspidi vaadeldavat saarte piirkonda, kus refräänsõnu kasutatakse regilauludes väga harva.

6 Andmestik, teema-analüüsi tulemused ning muud siinse uurimusega seotud failid on huvilisele kättesaadavad ja kasutatavad Eesti Keeleressursside Keskuse repositooriumi Entu vahendusel (Sarv 2020).

reetsele laulutüübile. Tabelis järjestasin teemad sisu alusel, paigutades lähestikku samadesse või lähedastesse laulurühmadesse kuuluvad teemad.

1	0,04	neidu peiu neiu pane tulid tõid saavanem sae siidisärki läks ehi kakku tulite mees aitimal pissi kuulis pange kannu vanema	pulm: ehtimine
2	0,04	tere venda neiu eile tuba tullid aitimal lähme andemasta peiu neitsikene metsa nurme tee korda tehtud neitsi õlut lauda uksi	pulm: teretamine
3	0,06	neiu neidu poiss tulla poisi ilus poissi õue aasi sööma tuleb viige kasvab viin kena tulge võtan näha neid nutab	pulm: viimine
4	0,06	pulma lehma kokku anda tegi natti palka härja vaka sukad anna püksid suure vei kivi ema kindad isa vaene nadu	pulm: veimed
5	0,04	äia ämma võta õde silmad pane meele anna vii nõuad tuleb nao mehe leina miele mine vennad küdi tubaje kera	pulm: õpetus
6	0,04	kiike kiige kiik metsa tehtud lähme poisid neidu poiste venna toodud aitimal pereisada tulge meeled põldu teinud neiud lauda pereemada	kiige
7	0,05	saani neidu küla neiu sai juua anna läks alli neiud tegi said isa kirju viisin vaene nägin ütles musta sõitsin	kosjad: saan
8	0,04	laeva neiud pani tuli riita valge vesi linna mehed kinni said teinud hakkas viru noored sadu pannud tulli udu vanad	kosjad: laev
9	0,04	ema isa vara tõuse tähte mari tütar vakka läksin mulla kuu muru mullu tulla tõusta valmistama kostis haua karja läks	ema haual
10	0,05	laula isa laulan ema kuku küla ohjad viisi laulu kurjad kinni kodu metsa ussi ütleb koju alba kardan moodi laulaksin	laula
11	0,05	läksin tuli emm läks ema isa lõid härjad andis eit kodu küsima ärjad piitsa nutad metsa koduje õue kuldine viis	korduslaul
12	0,11	saab vana sai palju minna suure lapsed mees naene kodu välja leiba saaks katki valla naese kätte lahti neid teeb	kodu tööd
13	0,04	läks karju lähme tehtud pea alli kulda tegi mere kulla irnu jäid hiie mehe õue venda saba poisid karva lahti	hobune
14	0,04	sai läks karu vanamees kivi vesi jähi otsas ätsed tamme kõvasi härg metsa must neitsit raius jäljed vana vee maale	karu
15	0,05	mõista kell piima kulda mere tegi kits metsa sõela heina pekki kuus tulnud lehti villa rauda oksa sepp õues õeke	kits
16	0,02	ann anne laeva annan koju saagu aita kodu läks isa hakkas tuleb anna laskem ema laevad korra ajaks purju vaske	lunastatav Ann
17	0,03	kaske pihta aita kargajada miku meeli läks eide kaski pelgajada läksin tuli kata piima pinu sinda tooma tooge peta mere	mehetapja Meeli
18	0,04	kassi sain sai saksad kass saksa linna kingad tegin lähme teeksin nina panin teista saab läksin olemine sadula tegemine mees	kass
19	0,08	kubjas mõisa läksin poisid isa tuli saksa linna viinad kodu õuna las kulla ema kasva kosjad hakkas valda pani kubja	kubjas
20	0,04	isa sõjas ema armas läks õue vene õde vend tuli poega meeste venda kallis venna kaasa vennad tunne korda läheb	nekrut

Tabel 1. Hiiumaa, Saaremaa ja Muhu 3673 regilaulutekstist rakenduse MALLET abil genereeritud 20 teemat, teemade Dirichlet-parameetrid, mis kajastavad vastava teema ligikaudset proportsiooni andmekogumis, võtmesõnad ja teema folkloristlikule sisule viitavad sildid.

Teemajaotuse võtmesõnadest on näha kõigepealt, et seekord eristuvad arvuti poolt genereeritud teemad üksteisest tõepoolest sisuliselt ning seostuvad regilaulu temaatiliste rühmade või konkreetsemate laulutüüpidega. Ühe teema võtmesõnade hulka koonduvad ka teema sisu seisukohalt oluliste sõnade vormid ja sünonüümid. Nii näiteks kuuluvad mõiste „ema" eri variandid, mis suuremas, põhjaeesti laulude rühmas jagunesid eri keelemurdeid esindavatesse teemadesse, siin üheskoos 11. teema alla (*emm, ema, eit*),[7] niisamuti leiduvad sõna *neiu* eri vormid viie, eelkõige pulma- või kosjalauludega seonduva teema võtmesõnade hulgas.

Teemade tegelik sisu on enamasti laiem, kui seda võtmesõnadest aimata on, hõlmavad ju 20 teemat vaadeldava lauluvara kogu selle mitmekesisuses. Kuigivõrd peegeldab siinne teemajaotus kindlasti ka eri laulutüüpide populaarsust, kuid laulutüüpide ja teemade vahekorra uurimiseks tuleks esmalt vaadeldava laulukogumi tüpoloogia üle vaadata ja tüübinimetused ühtlustada.

Järgnevalt kirjeldan lühidalt eri teemade sisu ja nendega seonduvaid laule, et selgitada arvutuslikult saadud sõnakogumite suhteid regilaulu liigilise ja tüpoloogilise jaotusega. Esimese viie teema (kokku *ca* 25 % laulukogumi mahust) võtmesõnad seostuvad üldisemalt pulmadega ja konkreetsemalt pulmarituaali erinevate etappidega.

1. pulmateemasse kuuluvad nende pulma episoodidega seotud laulud, mis on seotud pulma teekäikudega, pulmamajja sisenemise või sealt lahkumisega ning neiu ehtimisega, lisaks pulmalauludele ka mõned haakuvate teemadega laulud või motiivid („Laevamäng", „Nukumäng" jm).

2. pulmateemasse kuuluvad pulmaõue, pulmamaja, pruudi ja pruudi riiete teretamine, kiitmine ja kirjeldamine, pulmasöögid ja -joogid, pruudikoju sissepalumine ja sealt äraminemine koos tänamistega ning peigmehekodus pruudi nägu rituaalselt varjava peakatte uju äravõtmine ja noorpaari magamapanemine, lisaks teretamised muudes lauludes ning mõned kosjalaulud.

3. pulmateema seondub ühelt poolt pruudi kodust kättesaamise, väljameelitamise ning äraviimise lauludega, teisalt aga kuuluvad siia ka pulma sõimu- ja laidulaulud (ning vihapidamise teema ka väljaspool pulmi). Lisaks pulmalauludele liitub sellesse teemasse ka üsna mahukas noorte abielueelseid suhteid ja seksuaalseid ihasid, aga ka soovimatut ligitikkumist kajastav laulude rühm ning „Ätsemäng".

4. pulmateema on seotud eelkõige veimede ja pruudile tehtavate kingitustega, lauldakse nii veimede tegemisest, vedamisest, jagamisest kui ka kiitmisest ja laitmisest. Loodetakse, et pruut saab peigmehe perekonnalt kingiks koduloomi. Lisaks

7 Lisaks sellele teemale kuulub *ema* veel kuues teemas võtmesõnade hulka, kuid neis pole sünonüümsed variandid esile tulnud.

pulmalauludele kuuluvad teemasse eelnevaga seotult mitmesugused käsitöö ja riietega seotud laulud, loomade ja karjandusega seotud laulud, pulma või peole kutsumata jätmise laulud, mõned vaeslapselaulud ning usside, madude, konnade ja sisalikega seotud laulud ja loitsud.

5. pulmateema sisaldab õpetusi peamiselt pruudile, kuidas uues kodus olla ja käituda nii pulmade ajal kui pärast seda, aga ka peigmehele, keda manitsetakse pruuti hästi hoidma ja mitte vägivalda tarvitama. Õpetuslauludega seostuvad pulma episoodidest tanutamine ja põlletamine ning noorpaari magamapanemine. Teemaga liituvad veel mõned muud hoiatus- ja õpetuslaulud, laulud elust isa- ja mehekodus, murelaulud ja „Leinamäng".

6. teemasse kuuluvad eelkõige kiigelaulumotiivid ning ka sellega seonduv kolme metsa, järve, põllu nägemise kirjeldus. Teine suurem teemadering seondub õlle, selle valmistamise ja joomisega, pidutsemisega laiemalt ning sealjuures ka tantsu ja pillimänguga. Ilmse noorte teemaga seondub ka seksuaalne „kella kergitamise" motiivistik. Teema juurde kuuluvad tänamislaulud kokale, õlletegijatele, härjatapjatele ja ka pererahvale, kes on lubanud oma maja pidutsemiseks. Pidustused, söömine, joomine ja tänamine on neis lauludes nii mõnigi kord seotud pulmaolukorraga.

7. teema seostub kosjaskäiku ja naise valimist kujutavate lauludega, siia teemaplokki kuuluvad lüroeepilised jutustused saani tegemisest ja hobuse hoidmisest, mis tihti liituvad kaevul kosimise ning kalmuneiu motiividega, ning teiseks saartel hästi tuntud musta naise saamise laulutüübiga. Nendega liituvad veel motiivid, kus noormees väljas kohatud neidusid või neiud uhkelt ringi sõitvat noormeest vaatavad ja imetlevad, ning noormehe naisevaliku kriteeriumid (või vastavad õpetussõnad kellegi teise poolt).

8. teema keskmes on laeva ehitamine ja kosjaskäik laevaga, aga ka muud laevasõidu teemad (kauba toomine, laevahukk, pruut põgeneb meritsi). Laevasõidu teemaga liituvad veel ilmastikunähtusi puudutavad laulud ja motiivid (tuule tuba, sadu rikub riided, vihma Virumaale saatmine), linnas poes käimisega seonduvad teemad (sealhulgas „Väravamäng") ning kosjateemaga eri piirkondade neidude kirjeldused, naisevaliku soovitused ning pulmadeta meheleminek ning lapsesaamine. Selle teemaga seondub ka üsna palju kohanimesid alates lähimatest randadest ja laevateekonna kirjeldustest kuni kaugete sadamateni.

9. teema põhiosa moodustavad mitmesugused vaeslapselaulud, mille keskmes vanemate haual nende appikutsumine pulmadeks ettevalmistusi tegema, samasse rühma kuulub ka surma sajatamine. Teise suurema rühma moodustavad lüroeepikasse kuuluvad taevaste kosilaste laulud ja suure kuuse ümberistutamise motiiv. Lisaks kuuluvad selle teema alla veel laulud neiu tantsimaminekust ja pärast pois-

tega magamisest, samuti naisevõtusoovi väljendav laulutüüp „Murran mullikaid" ning piibu ja tubaka teemalised laulud.

10. teema sisaldab laia ja mitmekesise rühma laule lauludest ja laulmisest, sealhulgas ka mõned otseselt laulmist ja häält kiitvad või laitvad pulma sõimulaulud. Lisaks neile kuuluvad siia teemasse ka laulud, kus vanemad oma last kas halvustavad ja/või jätavad loomade-lindude hoida, ning väike rühm petja poisi või peiu laule, millest mõned kirjeldavad ka abielumehe vägivaldset käitumist.

11. teema sisaldab mitmesuguseid korduslaulude hulka liigituvaid laule – need on laulud, kus noor neiu või poiss satub täbarasse olukorda, kas on kadunud haned, hobune või kari, varastatud ehted, rikutud riided, kas on neiu langenud poiste rünnaku ohvriks või ei saa noormees omale sobivat naist. Pärast oma loo jutustamist vanematele saab õnnetu laps neilt abi või lohutust. Samasse teemasse kuulub veel teisigi lüroeepilisi süžeid (kadunud (pea)hari, tedrelaskja, karja kauplejad) ning neis lauludes ettetulevad motiivid ka iseseisvalt (neiu pole püssil püütav). Lisaks jutustavatele lauludele kuuluvad siia teemasse veel laulud, mis kirjeldavad, kui hea oli kodus kasvada (ja mõnel juhul tuuakse paralleeliks ka vastupidine olukord kas mehe kodus või võõrsil teenimas).

12. teema keskmes on kodused tööd, naised ja mehed. Suur osa lauludest on seotud noormehe kosjaminekuga ja kaalutluste või kirjeldustega, milline naine paremini tööd teeb. Siia liituvad ka naise ja neiu elukorralduse võrdlused. Kosjateema kõrval tulevad kõneks noorte abielueelsed seksuaalsuhted. Lisaks sellele seostub teema ka töödega mõisas, kurtmisega raske mõisatöö üle ja lootusega teotöölt pääseda. Teemale on iseloomulikud uuemasse regilaulu stiilikihistusse kuuluvad pikemad loetelud, kus (erinevalt tavapärasest regivärsi stiilist, mis otsest kordamist väldib) korratakse põhilisemat teemaelementi värsist värssi (kümme kosilast, nädalapäevade toimetused, loendid asjadest, mida on palju, naise kehaosade kirjeldus, erisuguste veskite kirjeldus jne). Veidi uuemas stiilis ja/või vähem arhailise keelega tekstidest kuuluvad siia ka osalt torupillilauludega segunenud tantsulaulud („Pill hüüab pinu taga"), mõned loitsud („Varesele valu") ja hulk siirdevormilisi või suisa uuemaid riimilisi laule, mis regilauluandmebaasi esialgses määratluses on ekslikult regilauluks klassifitseeritud (andmebaasitekstide aluseks olnud masinakirjakoopial puudus täpsustav märge laulu žanri kohta). Tegemist on mahu poolest kõige suurema teemarühmaga laulukogumis, kuid suur osa selle teema alla kuuluvatest sõnadest on jaotatud hajusalt laiali eri laulude vahel ning need koonduvad suhteliselt harvemini kompaktseteks terviklauludeks.

13. teema peategelaseks on hobune. Siia kuuluvad nii jutustavad, hobuse otsimisest ja hobusega kosjaskäimisest rääkivad laulud kui lühemad episoodid uhkest

hobusest ja tema seljas sõitjast ning ka „Hobusemäng". Metsast hobuse otsimise lauluga seonduvad ka muud metsas aset leidvad sündmused, sh pikemad jutustavad laulud hiies käimisest. Teema on valdavalt meestekeskne ja siia kuuluvad ka laulud ja motiivid ainsa venna tegemistest, naise otsimisest või kuldnaise tegemisest ning ei puudu ka laulud noormeeste ihast neidude järele ning naissuguelundist, kes omatahtsi mööda ilma ringi uitab. Teemasse mahuvad ka pikemad arendused käsikivist ja tema saamisest (ning hobustega vedamisest) ja ka üks pulma sõimulaulu motiiv (suust ja habemest tuleb suitsu).

Edasi järgnevad kaks ahellauludega seotud teemat. Ahellaulud koosnevad suuresti küsimuste ja vastuste jadadest, need on paljuski kuulunud laste repertuaari ning moodustavad koos mõnede teiste lastelauludega eraldi poeetilise süsteemi, mis on küll regilaulule väga lähedane ja võib sellega kergesti põimuda, kuid kasutab rõhulist värssi ning ka regilaulule iseloomuliku parallelismi kasutus on juhuslikum.

14. teema keskmes on nn karu ahel (mis tihti algab sõnadega *liiri-lõõri lõoke*), ning üks haruldasem, kuid karu ahelaga väga sarnane ahellaul (algab Jaani kaeralõikusega). Karu ahelale liitub sageli, kuid esineb ka iseseisva süžeena arendus karu muule maale minekust. Lisaks ahelatele liitub teemale muidki lastelaule, mis oma vormilt jäävad kõik regilaulu piirialadele, ning muudest lauludest „Sirbiviskamise laul". Regilaulupärasematest tekstidest kuuluvad siia mõned laulud, kus on tegemist loomade loeteludega, üks neiu ja noormehe heinamaal kohtumise episood ning „Laevamäng" koos imemaa kirjeldamisega. Siia teemasse kuulub ka pulmasõimulauludest üks loomadega seotud motiiv (laugi lehmaga ja kirju koeraga võitlemine).

Ka **15. teemaga** kõige selgemini seostuv laul on ahellaul, nimelt kitsest, kes heinu tooma läheb, kuid võrreldes eelmisega on teema ülejäänud sisu palju regilaululisem. Lisaks kitse ahelale kuuluvad peaaegu tervenisti sellesse teemasse veel ahelalaadsed mõistatuslaulud ning samuti küsimustel ja vastustel põhinev mütoloogiline laul merest meie õue all. See laul läheb nii mõnigi kord üle „Loomislauluks", millele lisaks seonduvad lüroeepilistest motiividest teemaga veel mere pühkimine ja venna otsimine. Ülejäänud selle teema laulud on üsna mitmekesised, siia kuuluvad mõned noorte suhteid käsitlevad regilauluteemad, hällilaulud ja mängitused, mõned loitsud, liisulugemised ning mängudest „Kuningamäng" ja „Kullimäng".

Järgneva kahe teema kõige iseloomulikumad esindajad on laulud, mida oma sisu ja stiili poolest võib seostada Lääne-Euroopa ballaaditraditsiooniga: laulud käsitlevad skandaalset sündmust, mis toimub nimeliselt tuntud isikuga (vt Asplund 1994, Kuusi 1963).

16. teema on teistega võrreldes väga kompaktne ning seondub kõige enam lauluga laeva tõmmatud või meelitatud ja väljalunastamist soovivast neiust, kelle nimeks saarte regilauluvariantides on Ann (riimilise laulu samateemalistes süžeedes on peategelaseks tihti Lilla). Teemasse kuulub veel lunastatava neiu lauluga sisuliselt seonduv ja motiiviti ka kattuv vanemate poolt müüdud neiu laul, laul Hollandi sulasest ning mõned sajatustekstid.

17. teema võtmelauluks on traagiline lugu mehetapjast, kes vaadeldava piirkonna lauludes kannab nime Meeli või Meelas. Teemasse kuuluvad veel lüroeepilised lood põllul kündva kure leidmisest ja kojuviimisest ning osalt ka mütoloogilised laulud neljast neiust ja mere pühkimisest. Lisaks lüroeepikale kuulub selle teema alla üsna mitmekesine ring laule, siia kuuluvad mardi- ja kadrilaulud, ahellaul metsa tuld tegema minevast väikeloomast (nn hiire ahel), kalapüügilaulud, laul noore ema hoolest, rannas ja metsas elamist võrdlevad laulud, mõned labasusse kalduvad laulud ja mitmeid loitsulaadseid tekste. 17. teema lauludes kasutatakse käskivat kõneviisi silmatorkavalt rohkem kui muudes teemades.

18. teema keskmes on saarte piirkonnas hästi tuntud ja levinud laul pahategijast kassist, kes laulu käigus kinni püütakse ja ära tapetakse. Lauluga liituvad mitmesugused eellood, kuidas tekib söök, mille kass nahka paneb, milleks võib olla kas suure härja või sea tapmine, haraka ja kiive (kiivitaja) tüli, kalalkäik, kitse toitmine vm. Need motiivid esinevad ka sama teema all iseseisvate lauludena. Teine teemasse kuuluv laulukogum seondub noormehe tegudega, keda vanemad ei luba kosja, ta teeb omale tuulest hobuse ja ehitab omale metsa maja, mida kõik imetlevad. Viimane motiiv võib esineda ka iseseisvalt ja sarnaneb samuti siia teemasse kuuluva lauluga neidude ehitatud linnast. Kolmas pikem teemasse kuuluv arendus seostub vaeslaste pisaratest kasvanud suure tammega, mille all voolavas jões püütud kaladest ja vähkidest saab põllurammu ja rikkust. Neljandaks alarühmaks on laulud sellest, mida laulik teeks, kui ta saaks (oleks minu olemine), peamisteks seonduvateks motiivideks on siin sakste töölepanek ja taevasse talli ehitamine.

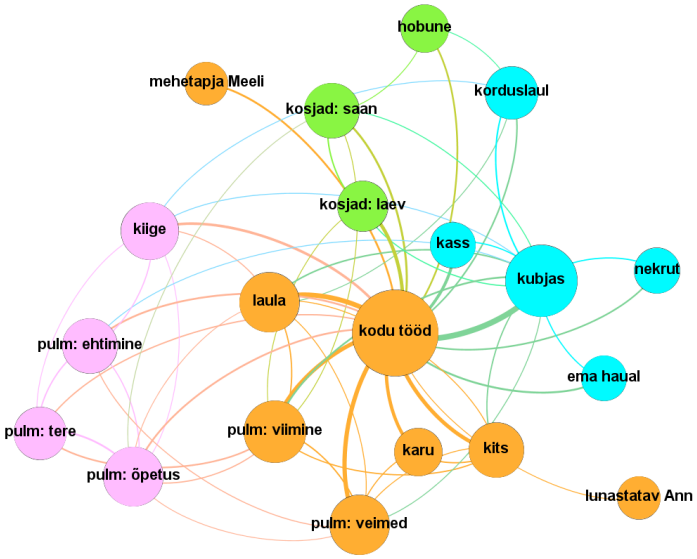
19. teema koosneb kahest suuremast alateemast. Esimeseks on mõisatööd ja tegemised, kurtmised mõisaorjuse üle ning vastuolud mõisa ametimeeste (eriti kupja) ja teoliste vahel, ka kupja sõimamine ja tapmine. Teise alateema keskmes on mitmesugused kosimisolukorrad neiu poolt vaadatuna (sageli kosilaste peletamine) ning sellega seonduvate olukordade kirjeldused (töötamine heinamaal, mäel tantsimine või metsas hulkumine, unemäele uinumine, neiu uhked ehted ja mis neist arvatakse). Lisaks kahele suuremale alateemale kuuluvad siia veel mõned väiksemad laulurühmad: laulud kündmisest ja viljalõikusest, laulud täide noppimisest, laulud

koduste halvast suhtumisest lapsesse, laulud noorest abielunaisest, kes töö juures ootamatult „põlvist põdema" jääb, laulud imelikust metskitsest ja imelisest kasukast.

Viimane, **20. teema** sisaldab laule nekrutiksvõtmisest ja sõjaskäimisest ning lisaks neile ka laule muidu kaugemalt koju pöördumisest tihti koos selgitustega, miks ja millal kodust lahkuti. Sellele küllalt kompaktsele laulurühmale lisaks kuulub teemasse veel väikelaste tammutamislaul „Kasva taadile tammeraiujaks".

Iga laul koosneb enamasti mitme teema elementidest, kõige olulisem teema moodustab laulust 19 % kuni 100 %, keskmiselt 68 %. Uurisin teemade omavahelisi seoseid selle alusel, kui sageli nad esinevad koos laulu kõige olulisema ja olulisuselt järgmise teemana.[8] Arvutasin kokku teemapaaride seosed ja kuvasin need võrgustikuna (kasutasin jällegi Gephit ja selle modulaarsusarvutust – joonis 2). Jooniselt näeme, et kesksel kohal on ka muidu kõige mahukam kodu ja koduste tööde teema, millel on seoseid kõige rohkemate teemadega. Kodulauludega samasse võrgustikujaotisse (mõtteliselt siis koduga seotud sfääri) kuuluvad veel osa pulmalaulu-teemasid, mõlemad ahellaulu-teemad, laulmise teema ja kaks naistekeskse lüroeepika teemat. Viimased on mõlemad küllalt selgepiirilised ja neil on vähem seoseid teiste teemadega. Lauludes, kus domineerib lunastatava Anne teema, on sellega seotud keskmiselt 76 % sõnadest, mehetapja Meeli teema puhul vastavalt 66 % (kõige madalam, 58 % on see näitaja kodu teema puhul, mis haakub kergesti kõigi teiste teemadega). Teemade seoseid kujutaval graafil näeme veel, et omaette koonduvad kokku pulmalaulud ja nendega lähedased, samuti kombestikuga seotud ja noorte suhteid puudutavad kiigelaulud. Eraldi võrgustikujaotusesse kuuluvad meestekesksed ja valdavalt meespeategelasega kosjalaulud ning hobuse ja hiie teema, moodustades justkui meeste iseseisvalt kodust väljas toimetamise sfääri. Omaette tegevussfääri kuuluvad mõisa ja sakstega seotud teemad, millega liituvad kaks jutustavate laulude teemat, mille tegevus toimub samuti pigem kodust eemal. Nii peegeldaks teemavõrgustik justkui sündmusi või tegevuspaiku, mille keskmes on kodu.

8 Selline mitme teema kooslus on mingis mõttes kõrvutatav folkloristikas kasutusel oleva kontaminatsiooni mõistega, kus üks laul võib koosneda mitmest laulutüübist või mitme laulutüübi elementidest. Erinevuseks on see, et siinse analüüsi teemad on määratletud sõnatasandiga ning eri teemad võivad laulus omavahel üsna segiläbi paikneda.



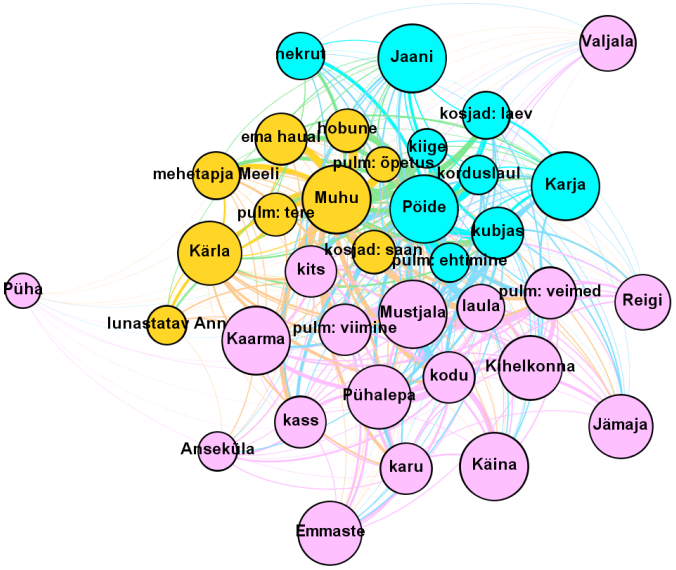
Joonis 2. Hiiumaa, Saaremaa ja Muhu lauludes esinevate teemade omavahelised seosed iga laulu kahe olulisema teema koosesinemiste põhjal. Joonisel on kuvatud seosed, mis esinevad rohkem kui 20 laulu puhul, sõlmede paigutus lähtub sõlmedevaheliste seoste arvust ja tihedusest – paigutuseks kasutasin Gephi algoritmi Force Atlas 2, joone jämedus kajastab teemade vaheliste seoste arvu, ringi suurus kajastab seostuvate teemade arvu, võrgustiku omavahel tihedamalt seotud alaosad on värvitud eri värvi.

Lisaks teemade eneste võrgustikule võimaldab andmestik ka vaadata, kuidas teemad jaotuvad kihelkonniti. Kihelkondade ja teemade omavahelised seosed on esitatud võrgustikuna joonisel 3. Graafikul kuvatud ringide suurus sõltub sellest, kui palju seoseid igal teemal kihelkondadega või kihelkonnal teemadega on. (Kuna kihelkondi on vähem kui teemasid, siis on ka teemade seostamispotentsiaal väiksem ja ringid üldiselt väiksemad kui kihelkondade omad.) Selle vaatluse kõige olulisemaks tulemuseks on, et teemad on jaotunud vaadeldava ala lõikes üsna ühtlaselt, üldiselt esinevad tuvastatud teemad kõigis kihelkondades, mis näitab taaskord, et siinse materjali puhul keeleline varieeruvus teemajaotuse genereerimisel suurt rolli mänginud ei ole. Võrgustik on sedavõrd tihe ja ühtlane, et ka võrgustiku jaotusanalüüs annab siin igal käivitamisel veidi erineva tulemuse, kordagi ei ole see aga vihjanud, et Hiiumaa laulud temaatiliselt (või keeleliselt) Saaremaa ja Muhu lauludest eristuksid.[9]

Võrgustikku silmitsedes võib siiski täheldada, et lüroeepiliste lauludega seonduvad teemad koonduvad pigem Muhu ja Pöide ümbrusse, kus regilaulutraditsioon on

9 Minu kogemuses on nii regilaulude meetrilise kui ka stilomeetrilise varieeruvuse alusel kihelkondade seosetiheduse põhjal joonistatud võrgustikud andnud tegeliku geograafiaga üsna hästi kooskõlas oleva tulemuse. Siinsel teemavõrgustikul on meie läänesaarte kihelkondade geograafia aimatav ainult tükati.

olnud rikkalikum ja mitmekülgsem, ning Hiiumaal ja idapoolsel Saaremaal domineerivad pigem laste- ja lüürikateemad. Pulmalaulude teemad asuvad pigem võrgustiku keskosas, seondudes enam-vähem ühtlaselt nii laulurohkete kui lauluvaesemate kihelkondadega.



Joonis 3. Hiiumaa, Saaremaa ja Muhu lauludes esinevate teemade seosed kihelkondadega iga laulu olulisema teema põhjal. Sõlmede paigutus joonisel lähtub sõlmede vaheliste seoste arvust ja tihedusest – paigutuseks kasutasin Gephi algoritmi Force Atlas 2, joone jämedus kajastab sõlmede vaheliste seoste arvu, ringi suurus kajastab vastavalt kas teemaga seostuvate kihelkondade või kihelkonnaga seostuvate teemade arvu, võrgustiku omavahel tihedamalt seotud alaosad on värvitud eri värvi.

Tulemuste arutelu

Keeleliselt võrdlemisi ühtlase ja lemmatiseerimata tekstikogumi, Hiiumaa, Saaremaa ja Muhu regilaulutekstide teema-analüüs võimaldas saada arvutuslike vahenditega kiire ülevaate küllalt suure hulga regilaulude peamistest teemadest, nende proportsioonidest, omavahelistest seostest ja piirkondlikust jaotumisest. Seejuures, nagu näitas arvutuslikult tuvastatud teemadesse kuuluvate laulude lähem vaatlus, on teemade sisu siiski mitmekülgsem, kui seda võtmesõnadest aimata võib. Siinses vaatluses, kus genereeriti 20 teemat (mis on üsna juhuslikult valitud arv), oli iga teema keskmes küll vaadeldavas laulukogumis oluline sisuvaldkond, kuid paljudel juhtudel võis eristada ka väiksemaid alateemasid, mille seos teema keskse sisuga polnud intuitiivselt alati päris selge (näiteks piibu ja tubaka laulud üldiselt vaeslapselauludega seonduvas teemas). Nii mõnelgi korral ilmnes, et sisult erinevate alatee-

made ühe teema alla kokku viimise aluseks olid kommunikatiivsed või laulu tegevusi raamistavad vormelid (näiteks palumine mardi- ja kadrilauludes ning mehetapjast jutustavas laulus või küsimus-vastus-struktuurid ahellaulude, mõistatuslaulude ja mõnede teiste laulutüüpide puhul) – mis juhib tähelepanu regilaulule iseloomulikule võttele kasutada poeetilisi vormeleid eri teema ja sisuga lauludes. Kui lähtuda selgete alateemade ilmnemisest arvuti poolt genereeritud teemade sees, siis tundub õige valida algoritmi rakendamiseks mõnevõrra suurem teemade arv. Tekstikogumi jaoks paraja teemade arvu kindlakstegemine on osutunud teema-analüüsi puhul pead murdma panevaks ülesandeks ning tegeletud on ka meetodite arendamisega, et see näitaja tekstide põhjal automaatselt tuvastada (nt Sbalchiero ja Eder 2020, Satpathy 2018). Mistahes arvu teemade puhul nõuaks analüüsitulemuste valiidsuse hindamine siiski ajamahukat arvukate tekstide sisusse süvenemist.

Uurimuse üheks lähteküsimuseks oli, kas ja kuidas mõjutab teemade modelleerimist eesti keelele iseloomulik sõnavormide paljusus lemmatiseerimata tekstide puhul. Ühest küljest võib öelda, et teemade tuvastamist ei seganud see kuidagi – arvuti oli võimeline tuvastama temaatiliselt kokkukuuluvad tekstid ja motiivid. Teisest küljest on aga ilmne, et lemmatiseerimata tekstide puhul osalesid teemade modelleerimisel ka grammatiliste vormidega edasiantavad sisu aspektid. Nii näiteks viitavad juba teemade võtmesõnade hulgas olevad tegusõnad oma vormi ja modaalsusega ka laulurühma žanrile ja funktsioonile: kombestikuga seotud laulude verbivormid on pigem olevikku või isegi tulevikku suunatud, jutustavates lauludes on tavalisemad minevikuvormid, õpetuslauludes levinumad sõnavormid on käskivas kõneviisis. Laulmise ja mõisateemaga seoses on oluliseks osutunud ka tingivas kõneviisis vormid, mis viitavad tegelikkuse erinevusele soovitavast: *laulaksin*, *teeksin*. Iseenesest on see regilaulude liigitamisel isegi hea valik, kuid siiski peab tulemuste interpreteerimisel meeles pidama, et tegemist ei ole grammatilistest aspektidest sõltumatu liigitusega. Laulukeele murdelise varieeruvuse küsimus õnnestus ilmselt laulukogumi piiramisega lahendada, uurimistulemused ei sisaldanud otseseid viiteid olulistele piirkondlikele erinevustele keeles ning teemad olid jaotunud kogu vaadeldaval alal võrdlemisi ühtlaselt.

Teiseks uurimisküsimuseks oli, kuivõrd mõjutavad laulutüüpide ebavõrdsed proportsioonid teema-analüüsi. Teemadesse kuuluvate laulude vaatlus näitas, et suurema üleskirjutuste arvuga laulutüübid ja ka korduvate vormelite või tekstiosadega laulutüübid (nagu meie vaatluses „Lunastatav neiu", mehetapjalaul, sõjalaul ning ka korduslaulu struktuuriga laulud) ning ka püsivama, vähem varieeruva tekstiga laulud (nagu meie vaatluses ahellaulud) satuvad sagedamini teemade keskmesse. Haruldasemad laulud ja laulurühmad kas moodustavad kõrvalteema mõne suurema teema all või pudenevad laiali eraldi teemade vahel. Kuigivõrd aitaks selle mure vastu kind-

lasti suurema arvu teemade genereerimine. Praegune laulukogumi jaotus 20 teemaks vastab oma mahult ligikaudu regilaulude folkloristlikule liigitusele temaatilis-funktsionaalseteks laulurühmadeks (vrd nt antoloogia „Eesti rahvalaulud", kus on 14 laulurühma, mis jagunevad omakorda alateemadeks ning kokku 3286 tüübiks). Vaadeldud kogumi jaotamine suuremaks arvuks teemadeks tundub mõistlik, arvestades mõnede siinses vaatluses tuvastatud teemade kirevat sisulist koostist, kuid samas kaotaksime analüüsitulemuste ülevaatlikkuses. Teoreetiliselt on võimalik ka lasta genereerida teemad ligikaudu laulutüübi mahus, kuid esialgu veel on läbi proovimata, kui tulemuslik selline jaotus oleks, kas ta oleks võimeline tuvastama teemadena samasse folkloristlikku tüüpi kuuluvaid laule või pudeneksid lalulutekstid praegusest veelgi enam eri teemade vahel laiali ning kuivõrd mõjutaksid tulemusi laulutüüpide erinevad esinemissagedused. Tegin ka katse laulukogumi jagamiseks 50 teema vahel ning selles jaotuse tulemused on mõnevõrra täpsemad ning mitmed puudused lahenenud, nii näiteks on eraldi teemadesse määratud mehetapjalaul ning kadri- ja mardilaulud. Kui 20 laulu puhul oli suurima teema osatähtsus laulus keskmiselt 68 % sõnadest, siis 50 teema puhul on see näitaja püsinud ligikaudu samas suurusjärgus, 64 %.

Kolmandaks uurimisküsimuseks oli, kuidas suhestub teema-analüüsi tulemusel saadud liigitus folkloristlike liigitustega. Teema-analüüs lähtub otseselt uurimiseks valitud kitsama ala materjalist ja selles leiduvate laulude sagedusjaotusest, mitte folkloristlikust tüpoloogiatraditsioonist, mis arvestab kogu Eesti ala lauludega ning osalt ka nende žanriliste ja vormiliste tunnuste ning funktsiooniga (vt Tedre 1974). Seetõttu toob see ka esile veidi teistsugused teemarühmad, kui oleme harjunud nägema regilaulude akadeemilistes väljaannetes. Erinevused pole mõistagi kardinaalsed, kuid näiteks kalendrilauludest tõusevad eraldi rühmana esile üksnes kiigelaulud. Samuti on teema-analüüs ühte teemasse kokku toonud folkloristlikus liigituses eri laululiikidesse kuuluvad, kuid sama teemaga seotud laulud (sellisteks teemadeks on siinses analüüsis näiteks laevasõit ja hobune, mis ühendavad vastavaid lüürilisi ja lüroeepilisi laule ning mänge). Mõnes kohas järgib aga siinne teema-analüüs regilaulu liigilist jaotust üllatavalt hästi, seda võib täheldada nii pulmalaulude puhul, mis tõesti näivad olevat koondunud kindlatesse teemadesse, kui ka näiteks lüroeepilistest lauludest perekonnaballaadide puhul, millest suurem osa on koondunud üheteistkümnendasse, korduslaulude teemasse.

Üheks põnevamaks tulevikuväljavaateks on uurida, kuivõrd teemad kajastavad regilaulu ajaloolisi või stilistilisi kihistusi. Nägime eespool, et omaette teemadesse olid koondunud laste- ja mängulaulud, mille meetriline eripära ei nõua kindlat silbiarvu ja seega pole neis lauludes (ka muudes Eesti piirkondades) nii suurt tarvidust säilitada arhailisi keelendeid. Samamoodi olid eraldi teemadesse koondunud teada-

olevalt hilisemasse kihistusse kuuluvad ja ka keeleliselt seetõttu uuemad nekrutilaulud ja mõisateemaga seonduvad laulud ning ühe teema alla olid koondunud ka laulukogumisse sattunud päris uuemasse kihistusse kuuluvad tantsulaulud, siirdevormilised ja riimilised laulud. Kas ja kuivõrd ülejäänud teemarühmad võiksid seonduda keele (ja seega ka laulude kujunemisloo) eri kihistustega, jääb edaspidise uurimise ülesandeks. Teadaolevalt eesti regilaul romantilisi tundeid eriti ei väljenda, seetõttu on mõneti märgiline, et sõnad *armas* ja *kallis* on võtmesõnadeks uuemate nekrutilaulude puhul, seevastu pulmalauludes olulisteks omadussõnadeks on hoopis *ilus* ja *kena*.

Kokkuvõtteks

Siinse artikli vaatlused regilaulude teema-analüüsi võimalikkuse ja mõttekuse osas võib kokku võtta järgmiselt: (1) teema-analüüsi jaoks on mõistlik kasutada võrdlemisi ühtlase keelekujuga materjale, vastasel korral hakkavad keelelised erinevused mõjutama teema-analüüsi tulemusi; (2) teema-analüüsiks on võimalik kasutada lemmatiseerimata tekste, kuid sellisel juhul osalevad teemade genereerimises ka grammatilised kategooriad (nt verbi aeg või kõneviis), mis võivad osundada tekstide laadile, žanrile ja funktsioonile; (3) varieeruvate ja püsivate elementide (laulutüübid, motiivid) proportsioonid ainekogumis mõjutavad selgesti teemade moodustamist: mida sagedamini element materjalis esineb ja mida püsivama sõnastusega see on, seda suurema tõenäosusega moodustab ta teema keskme, samal ajal jäävad harvaesinevad sisuelemendid tähelepanu keskmest välja ja pigem kipuvad jagunema eri teemade vahel; (4) ühe teema alla koondunud sõnakogumid võivad, lisaks samale sisuvaldkonnale, viidata ka laulu sisu raamistavatele elementidele, näiteks ümbrusele (mets), käitumis- või suhtlusmallidele (näiteks millegi/kellegi palumine) ja vormellikele elementidele lauludes.

Võrreldes regilaulude folkloristliku liigitusega (1) tõstab automaatne analüüs rohkem esile tekstikogumis sagedamini esinevat sisuainest, tüüpe ja motiive (nii näiteks tõstab eraldi teema fookusse kiigelaulud, erinevalt folkloristlikust liigitusest, mille üheks peakategooriaks on kalendrilaulud); (2) jätab automaatne analüüs osaliselt märkamata žanripiirid (nii näiteks leiab teema-analüüsis mängud eri teemade alt, folkloristlikus liigituses moodustavad need aga eraldi laulurühma).

Allikad

Aavik, Johannes. 1914. *Eesti rahvusliku suurteose keel.* Keelelise Uuenduse Kirjastik 1. Tartu: Reform.

Asplund, Anneli. 1994. *Balladeja ja arkkiveisuja: Suomalaisia kertomalauluja*. Suomalaisen kirjallisuuden seuran toimituksia 563. Helsinki: Suomalaisen Kirjallisuuden Seura.

Barabási, Albert-László. 2016. *Network Science* (veebileht). http://networksciencebook.com/.

Bastian Mathieu, Sébastien Heymann ja Mathieu Jacomy. 2009. „Gephi: an open source software for exploring and manipulating networks." – *International AAAI Conference on Weblogs and Social Media* 8, 361–362. San Jose, California, USA.

Blei, David M., Andrew Y. Ng, Michael I. Jordan ja John Lafferty. 2003. „Latent Dirichlet allocation." – *Journal of Machine Learning Research*, nr 3, 993–1022. https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf.

Blondel, Vincent D., Jean-Loup Guillaume, Renaud Lambiotte ja Etienne Lefebvre. 2008. „Fast unfolding of communities in large networks." – *Journal of Statistical Mechanics: Theory and Experiment*, nr 10, P10008. https://iopscience.iop.org/article/10.1088/1742-5468/2008/10/P10008.

Eder, Maciej, Mike Kestemont ja Jan Rybicki. 2013. „Stylometry with R: a suite of tools." – *Digital Humanities 2013: Conference Abstracts,* 487–489. University of Nebraska-Lincoln, NE

Eilat, Toomas. 2019. „Tõde ja Õigus IV osa topic modelling." – *Toomase blogi.* 15. juuli. http://eilat.ee/2019-07-15-tode-ja-oigus-topic-modelling/.

Eesti rahvalaulud. Antoloogia. Toimetanud Ülo Tedre. Tallinn: Eesti Raamat 1969–1974. http://haldjas.folklore.ee/laulud/erla/.

Graham, Shawn, Scott Weingart ja Ian Milligan. 2012. „Getting Started with Topic Modeling and MALLET." – *The Programming Historian* 1. https://doi.org/10.46430/phen0017.

Ivanova, Maria. 2020. „What happened to Lenore? Describing changes in ballad genre by means of a statistical model." – *AKK Sügiskool: Ajaloo, kirjanduse ja kultuuriteaduste tudengikonverents. Teesid*, 9–10. https://sisu.ut.ee/sites/default/files/akksygiskool/files/teesid_theses_0.pdf.

Krikmann, Arvo. 1997. *Sissevaateid folkloori lühivormidesse* I. *Põhimõisteid. Žanrisuhteid. Üldprobleeme*. Tartu: Tartu Ülikooli Kirjastus. http:// folklore.ee/~kriku/LEX/KATUS.HTM.

Kuusi, Matti. 1963. *Suomen kirjallisuus* I. *Kirjoittamaton kirjallisuus.* Helsinki: SKS & Otava.

Lindström, Liina. 2015. *Ülevaade eesti murrete korpusest*. Tartu. https://www.keel.ut.ee/sites/default/files/fl/emk_teejuht2015.pdf.

McCallum, Andrew Kachites. 2002. *MALLET: A Machine Learning for Language Toolkit*, versioon 2.0.8. Alla laetud: 10. märts 2019. http://mallet.cs.umass.edu.

Mäkelä, Eetu. 2018. „Digging into a method: topic modeling" – *Computational literacy for the humanities and social sciences.* https://jiemakel.gitbook.io/clit4hss/computational-data-analysis-method-literacy/digging-into-a-method-topic-modeling.

Mölder, Martin 2019. „Valimisprogrammide kvantitatiivne tekstianalüüs." – *Martin Mölder* (blogi). 18. veebruar. https://martinmolder.com/blog/valimisprogrammid2019/.

Oras, Janika, Liina Saarlo ja Mari Sarv. 2003–2020. *Eesti regilaulude andmebaas*. Tartu: Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiiv. http://www.folklore.ee/regilaul/. https://doi.org/10.15155/9-00-0000-0000-0000-0008FL.

Peegel, Juhan. [1954] 2006. *Eesti vanade rahvalaulude keel*. Tallinn: Eesti Keele Sihtasutus.

Sarv, Mari. 2015. „Regional Variation in Folkloric Meter: The Case of Estonian Runosong." – *RMN Newsletter*, 9: 6–17.

———. 2019. „Poetic metre as a function of language: linguistic grounds for metrical variation in Estonian runosong." – *Studia Metrica et Poetica* 6 (2): 102–148. https://ojs.utlib.ee/index.php/smp/article/view/smp.2019.6.2.04.

———. 2020. „Hiiumaa, Saaremaa ja Muhu regilaulude teema-analüüsi failid." – *ENTU. Eesti Keeleressursside Keskus.* https://entu.keeleressursid.ee/entity/document/9771.

Satpathy, Amit. 2018. „Topic Modeling with Automated Determination of the Number of Topics." – *GitHub.* https://github.com/bademiya21/Topic-Modeling-with-Automated-Determination-of-the-Number-of-Topics/commits?author=bademiya21.

Sbalchiero, Stefano ja Maciej Eder. 2020. „Topic modeling, long texts and the best number of topics. Some Problems and solutions." – *Qual Quant*, nr 54, 1095–1108. https://doi.org/10.1007/s11135-020-00976-w.

Tedre, Ülo. 1974. „Rahvaluule ja rahvalaul." – *Eesti rahvalaulud. Antoloogia.* IV. Toimetaja Ülo Tedre, 9–19. Tallinn: Eesti Raamat.

———. 1998. „Rahvalaulud." – *Eesti rahvakultuur*, toimetanud Toomas Tamla, 548–564. Tallinn: Eesti Entsüklopeediakirjastus.

„Topic Analysis." – *MonkeyLearn* (veebileht). Vaadatud: 14. detsember 2020. https://monkeylearn.com/topic-analysis/.

Uiboaed, Kristel 2018. „Eestikeelsete stoppsõnade loend." – *Tekstikaeve* (blogi). 18. aprill. http://www.tekstikaeve.ee/blog/2018-04-18-eestikeelsete-stoppsonade-loend/.

Mari Sarv – PhD, Eesti Rahvaluule Arhiivi vanemteadur. Tema peamiseks uurimisteemaks on eesti regilaul. Sel teemal on ta avaldanud kaks monograafiat (2000, 2008), korraldanud konverentse, toimetanud artiklikogumikke, samuti osalenud regilaulude andmebaasiga seotud töödes. Lisaks regilaulu-uurimisele on ta korraldanud välitöid, pannud aluse Eesti digitaalhumanitaaria konverentside seeriale ning osalenud Eesti Kirjandusmuuseumi digitaalarhiivi rajamisel.
E-post: mari[at]haldjas.folklore.ee

SUMMARY

Topic Analysis of Estonian Runosongs: Prospects and Challenges

Mari Sarv

Keywords: runosong, topic modelling, corpus-based analysis, folklore, digital humanities

The article explores possibilities of computational topic analysis of Estonian runosong texts using the latent Dirichlet allocation (LDA) topic modelling. Runosong is an oral poetic tradition known among most of Finnic peoples. Estonian runosong texts, the material of the current research, have been collected mainly since 1880s and gathered into the Estonian Folklore Archives of the Estonian Literary Museum, where the runosong database with more than 100 000 texts has been compiled (Oras et al 2003–2020). Language of runosongs varies considerably across dialects and, in addition to that, it uses a specific archaic idiom different from the spoken language which complicates the computational analysis of the content aspects of the texts.

Topic modelling is a method that enables to discover abstract topics detected statistically on the basis of the frequency of the co-occurrence of the words in the texts. In case of a runosong corpus, the method could be used to automatically detect the thematic structure of a large amount of runosong texts, to compare the thematic distribution of regional traditions of the runosong, and to analyse how the thematic distribution obtained with the help of computational methods relates to the classification of the texts resulting from folkloristic analysis. The idea of the current article is to explore whether topic modelling can give meaningful results if applied to unlemmatized and highly variative runosong texts.

For LDA topic modelling I used the application MALLET (McCallum 2002). The initial trials with the whole corpus of runosong texts made it clear that the language of the songs is too variative to reach the level of content. It also became obvious that it is necessary to remove stopwords and refrain words. The topics, obtained from the runosongs from all over Estonia, represented dialectal variants of the language rather than thematic clusters and it was necessary to restrict the material. I used stylometric analysis (using R package stylo, Eder et al 2013) to divide the area into linguistically more homogenous subregions, and chose the area of Western islands of Estonia with 16 parishes and 3672 song texts for further explorations.

With this material I decided to generate 20 topics. Within this smaller area the topics did not cluster regional language variants any more: (1) the linguistic variants of the main concepts of a topic were brought together under the keywords of the same topic; (2) in most cases, the detected topics were distributed among all the parishes included in the selection.

Looking at the 20 keywords, the topics indeed seemed to reflect certain thematic subgroups of the songs. In several cases the most prominent song type of a topic was reflected in keywords, in other cases the keywords referred to larger groups of songs. Five of the 20 topics focused on weddings, more precisely, on different episodes of the wedding ritual: adornation and dressing, arriving and greeting, finding the bride and taking her to her new home, sharing the presents prepared by the bride, and recommendations to the bride and the groom. In all these topics the verbs refer either to the present or the future (rather than to the past which is common in narrative songs). A topic of swinging songs includes also the songs about dancing and feasts. Five topics focus on different narrative plots about the troubles of young people, about wooing and marriage. Lyric songs about the life of orphans and about singing form a separate topic each,

and there is a separate male topic covering the songs of various genres related to horses, riding and the woods. The largest topic includes the songs on working at home and outside, but also the songs about premarital sex. There are two topics with the focus on well-known children's songs and lullabies. Two topics relate to German landlords, their power and activities, and one to recruiting and the war.

As a conclusion of this exploration: (1) for topic modelling it is necessary to use the texts in homogenous language variants; otherwise, the linguistic differences override the topics at some point; (2) it is possible to use unlemmatized texts for topic modelling, but in this case the grammatical features (tense, modality) interfere with topic analysis; (3) the proportions of variable and stable (recurrent) elements (song types, motifs) in the material have a clear impact on topic formation: the more frequently an element occurs in the material, and the more stable is its wording, the bigger its probability to form the centre of a topic, whereas distinct but rare themes remain unnoticed and will be shared between the topics of more prominent subjects; (4) common sets of words assembled together as the topic may, in addition to the common thematic focus, refer to a common framework, for example environments, and behavioural or communicative patterns (for example, begging for something). Compared to the folkloristic classification of folk songs, the automatic distribution of songs (1) highlights the subjects occurring more frequently in the body of songs (for example, a topic highlights swinging songs instead of calendar songs of the folkloristic classification); (2) partly overrides the genre differences (for example song games can be found under different topics, whereas forming a distinct group in folkloristic classifications).

Mari Sarv – PhD, senior researcher at the Estonian Folklore Archives (Estonian Literary Museum). Her main subject of study is Estonian runosong tradition. She has published two monographs on the topic, organized the conferences, edited the proceedings, and participated in compiling and developing the database of Estonian runosongs. She has also organized several folkloristic fieldworks, has initiated Estonian DH conference series, and has contributed to the establishing of the digital archival system of the Estonian Literary Museum.
E-mail: mari[at]haldjas.folklore.ee