E L O R E

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Artikkelit

https://doi.org/10.30666/elore.126008

Eteneminen omalla vastuulla

Lähdekriittinen laskennallinen näkökulma sähköisiin kansanrunoaineistoihin

Kati Kallio, Maciej Janicki, Eetu Mäkelä, Jukka Saarinen, Liina Saarlo ja Mari Sarv

Historiallisista, Suomen ja Viron kansallisvaltioiden syntyyn liittyvistä syistä johtuen itämerensuomalaista *runolaulua* eli niin kutsuttua kalevalamittaista kansanrunoutta (viroksi *regilaul*) on tallennettu, arkistoitu, järjestetty ja digitoitu poikkeuksellisen paljon. Kaiken kaikkiaan tällä hetkellä on sähköisenä käytettävissä 283 206 pääosin kalevalamittaiseen suulliseen runouteen liittyvää tekstiä viron (sis. seto), karjalan, lyydin, inkeroisen, vatjan ja suomen kielillä. Avoimesti verkossa olevat korpukset ovat monin osin syrjäyttäneet painetut julkaisut, arkistojen kortistot ja alkuperäiskäsikirjoitukset tutkimuksen lähteinä.[1]

Sähköiset ja laskennallisetkin lähestymistavat ovat mahdollisia, koska käytettävissä on määrällisesti suuri, riittävän jäsentynyt, pitkän tekstualisaatio- ja järjestelyprosessin tuloksena syntynyt aineisto, jota on myös tutkimuksessa tarkasteltu monenlaisista eri suunnista. Manuaalinen tutkimus on usein rajannut aineistoa esimerkiksi runotyypin tai -lajin, laulajan, suvun, pitäjän, alueen tai tallentajan mukaan ja tuottanut näkökulmia, jotka ovat välttämättömiä myös isompia aineistokokonaisuuksia käsittelevien laskennallisten tulosten tulkinnassa. Laskennallisella folkloristiikalla ja sähköisten kansanperinneaineistojen tutkimisella puolestaan on paitsi runsaasti tuoreita sovelluksia, myös pitkä historia (ks. esim. Vikis-Freibergs ja Freibergs 1978; Rüütel ja Haugas 1990; Harvilahti 1992; Sarv 2008; Tangherlini 2016; Ilyefalvi 2018; Harvilahti 2019).

Aineiston ominaispiirteet asettavat reunaehdot sähköisilläkin välineillä tehtäville tulkinnoille. Itämerensuomalaisen kansanrunousaineiston käyttäminen edellyttää melko laajaa kuvaa sen syntyhistoriasta, luonteesta, painotuksista ja epätasaisuuksista. Aineisto on sekä määrällisesti että laadullisesti laaja ja varioiva. Jo yksin kieliasun variaatioon vaikuttavat moninaiset aineiston sisältämien itämerensuomalaisten kielten murteisiin, itse runokieleen, sanojen taivutusmuotoihin ja tallentajien käyttämiin tallennusmenetelmiin ja erilaisiin

1 Artikkeli on tehty osana Suomen Akatemian hankkeita 333138 ja 346342, Viron tiedeneuvoston (Eesti Teadusagentuur) hanketta PRG1288 sekä Viron opetus- ja tiedeministeriön hanketta EKKD 126. Tekstin taustalla on kirjoittajien vuodesta 2019 lähtien FILTER-hankkeen (2020–2024, *Formulaic intertextuality, thematic networks and poetic variation across regional cultures of Finnic oral poetry*) piirissä tekemä tiivis yhteistyö ja aiemmatkin keskustelut kollegoiden kanssa. Erityisesti olemme kiitollisia Senni Timoselle, Janika Orakselle, Lauri Harvilahdelle, Frogille sekä kahdelle anonyymille vertaisarvioijalle. FILTER on Suomen Akatemian konsortiohanke SKS:ssa ja Helsingin yliopistossa (HELDIG), yhteistyössä Viron Kansanrunousarkiston kanssa. Kati Kallio on kirjoittanut artikkelin tekstin, muut tekijät ovat lisänneet tekstiin lyhyempiä jaksoja, tietoja, kuvaajia ja kommentteja: tästä johtuu myös artikkelin painottuminen suomalaiseen aineistoon.

kirjoitusasuihin liittyvät tekijät. Koko aineiston kattavia sanakirjoja, hakuteoksia tai kielellistä jäsennystä tekeviä parsereita ei ole, ja toisaalta aineisto on variaatioonsa nähden liian pieni laskennallisten kielimallien kouluttamiseen. Niinpä tietyt laskennalliset menetelmät kuten esimerkiksi tekstin automaattinen morfologinen jäsennys tai sanojen normalisointi eli lemmaaminen eivät ole kokonaisaineistoon sovellettavissa – joskus ollenkaan, joskus ilman huomattavaa työtä joko aineiston esikäsittelyssä tai menetelmäkehityksessä. Tämä tarkoittaa, että myöskään pidemmälle menevät sisällön laskennalliset analyysit kuten aihemallinnus (*topic modelling*) eivät tällä hetkellä ole sovellettavissa monikieliseen kokonaisaineistoon (soveltamisesta yhteen kielimuotoon ja kieliasultaan normalisoituun virolaiseen ERAB-korpukseen ks. Sarv 2020).

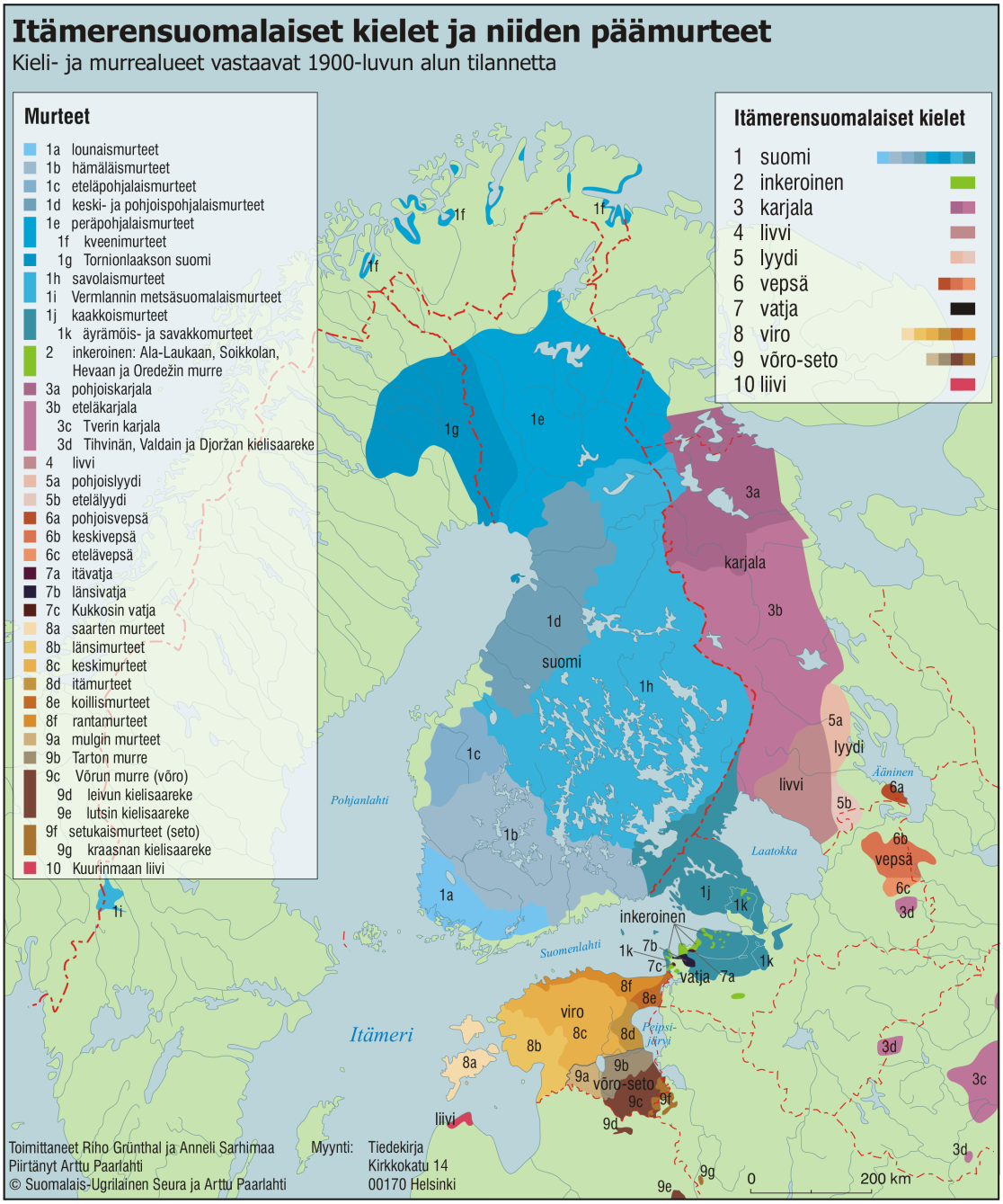
Sähköisten aineistojen ja laskennallisten välineiden käyttäminen edellyttää monipuolisia lähdekriittisiä otteita. Osin nämä liittyvät kansanrunoaineiston suureen kielelliseen ja poeettiseen variaatioon ja edustavan osa-aineiston tavoittamisen ongelmiin vaikkapa sanahakujen tai samankaltaisuuslaskentojen avulla (ks. Kallio ja Mäkelä 2019; Kallio ym. 2022). Osin kyse on aineiston pitkän tallentamis- ja järjestelyhistorian tuottamista painotuksista ja esimerkiksi metatietojen puutteellisuuksista.

Tässä artikkelissa analysoimme suomalaisen ja virolaisen sähköisessä muodossa olevan runolauluaineiston määrällisiä perusominaisuuksia, jotka liittyvät läheisesti aineiston pitkään syntyhistoriaan. Luomme ensin yleiskuvaa aineistojen järjestely- ja digitointihistoriasta ja kerromme, miten olemme FILTER-hankkeessa käsitelleet sähköisiä korpuksia. Sitten tarkastelemme aineiston ja tallennushistorian luonteeseen liittyviä määrällisiä ominaisuuksia ja lopuksi työn aikana paljastuneita aineiston ja sen metatietojen ongelmakohtia, joista on hyvä olla tietoinen myös laadullista tutkimusta tehtäessä. Tarkemmat analyysit esimerkiksi monista tallennushistorian yksityiskohdista, runotyyppihakemistoista, aineiston runo- ja säetason samankaltaisuuksista sekä suullisten ja kirjallisten runojen suhteista joudumme rajaamaan tämän artikkelin ulkopuolelle. Aineistona käytämme kahta suomalaista (SKVR, JR) ja yhtä virolaista (ERAB) sähköistä korpusta.

Virolaiset, venäjänkarjalaiset ja suomalaiset kansanrunokokoelmat

Kalevalamittaista runoutta tiedetään käytetyn suurimmalla osalla itämerensuomalaisista kielistä: ainoastaan vepsän ja liivin alueilta kalevalamittaista runoutta ei tunneta (runomitan variaatioista ks. Sarv 2019; Sarv ym. 2021; itämerensuomalaisista kielistä esim. Bakró-Nagy ym. 2022; Abondolo ja Valijärvi 2023; ks. Kuva 1). Aiemmat käsitykset kielistä ja kansakunnista näkyvät yhä nimityksissä: esimerkiksi 1900-luvun alun kirjasarja *Suomen Kansan Vanhat Runot* eli *SKVR* (1908–1948, 1997) pitää nimestään huolimatta sisällään paitsi suomen- myös karjalan-, inkeroisen- ja vatjankielisiltä tallennettuja runoja, joiden esittäjistä monet eivät kokeneet olevansa suomalaisia tai kuuluneet samaan hallinnolliseen kokonaisuuteen kuin nykyinen Suomen alue. (Itämerensuomalaisen kansanrunouden nimityksistä esim. Rüütel 1999; Kallio ym. 2017; Kalevalasta, kansanrunoudesta ja omistajuuksista ks. Tarkka ym. 2019.)

Käytännön kokoelmien tasolla itämerensuomalaisesta runolaulusta on tallentunut kolme melko erillistä kansallista kokonaisuutta: virolainen, venäjänkarjalainen ja suomalainen. Näitä on käsitelty pääosin erillisissä, joskin vuorovaikutteisissa tutkimustraditioissa. Jaon voi karkealla tasolla hahmottaa paitsi kansallisena ja valtionrajojen määrittämänä, myös kielellisenä: eteläiset itämerensuomalaiset kielet eli pohjois- ja eteläviro sekä seto löytyvät



Kuva 1. Itämerensuomalaiset kielet ja niiden päämurteet 1900-luvun alussa. Monikieliset tai -etniset alueet, substraatit ja kielihistorialliset suhteet eivät kartassa näy. Kuva: Toimittaneet Riho Grünthal ja Anneli Sarhimaa, piirtänyt Arttu Parlahti 2004/2012. Suomalais-Ugrilainen Seura. https://www.sgr.fi/fi/items/show/602.

pääosin virolaisista kokoelmista, keskenään läheiset pohjoiset itämerensuomalaiset kielet karjala, lyydi, inkeroinen ja suomi pääosin suomalaisista ja venäjänkarjalaisista kokoelmista. Virolle kielihistoriallisesti läheisen vatjan kielen vanhimmat kokoelmat on tallennettu Suomeen ja julkaistu SKVR-sarjan osana, 1900-luvun tallenteista taas suurin osa on virolaisten tutkijoiden tallentamia ja suurelta osin ääniteaineistoa (ks. Ariste 1960). Inkerinsuomalaisilta ja inkeroisilta 1900-luvulla tallennettua aineistoa sisältyy kaikkiin kolmeen kokonaisuuteen, neuvostoaikaista erityisesti virolaiseen ja venäjänkarjalaiseen.

Suurin osa runolaulun vanhimmista suomen-, karjalan-, inkeroisen- ja vatjankielisistä kokoelmista sijaitsee Suomalaisen Kirjallisuuden Seuran (SKS) arkistossa. Vanhimmat ja laajimmat vironkieliset aineistot sijaitsevat Viron Kirjallisuusmuseon (Eesti Kirjandusmuuseum, KM) Kansanrunousarkistossa (Eesti Rahvaluule Arhiiv, ERA). Etenkin 1900-luvun aineistoja on runsaasti myös monissa muissa arkistoissa ja instituutioissa.

1900-luvun karjalankielistä aineistoa löytyy etenkin Petroskoin Karjalan tutkimuskeskuksessa sijaitsevan Kielen, historian ja kirjallisuuden instituutin arkistossa. Kokoelmasta suuri osa on käsikirjoituksia tai äänitteitä, mutta aineistoa on myös julkaistu kirjamuodossa (esim. Jevsejev 1950; 1994; Lavonen 1989; Stepanova 2000; Mironova 2006; ks. myös Kuusi 1977) ja etenkin äänitteitä digitoitu. Arkistossa on paljon aineistoa myös aiemmin niukalti dokumentoiduilta pohjoisimmilta ja itäisimmiltä karjalankielisiltä alueilta (ks. esim. Lavonen 1989; Mironova 2006). Instituutissa on tekeillä runotekstien tietokanta, mutta suurin osa runoista ei ole vielä saatavilla sähköisessä muodossa (Kundozerova 2022).

Tässä artikkelissa tarkastellaan ERAB-, SKVR- ja JR-korpuksiin sisältyvää aineistoa (ks. Kuva 2). Vuodesta 2004 lähtien on avoimesti verkossa käytettävissä ollut suomalainen *SKVR*-tietokanta, joka sisältää *Suomen Kansan Vanhat Runot*-kirjasarjassa aikanaan julkaistut runot – koko aineisto saatiin tietokantaan vuonna 2006. Virolainen runolaulun tietokanta *Regilaulude andmebaas* (ERAB) avattiin verkossa vuonna 2010 ja täydentyy yhä. Tietokantojen käyttöliittymissä aineistoa voi etsiä sanahakujen, tallentajien, vuosien, paikkakuntien tai runotyyppihakemistojen avulla. (Laaksonen ja Saarinen 2004; Saarinen 2006; Harvilahti 2013; Järv 2016; Sarv ja Oras 2020.) Lisäksi SKS:n julkaisemattomien runojen kopiokortisto on digitoitu tekstimuotoiseksi JR-korpukseksi. Kokonaisaineisto on suuri, mutta esimerkiksi vanhimmat 1600-luvun aineistot, nuottien yhteydessä esiintyvät runot ja 1900-luvun äänniteaineistot sekä sananlaskut ja arvoitukset eivät sisälly näihin korpuksiin kattavasti.

Aineisto	**ERAB (2023)**	**SKVR**	**JR**
Tekstejä	108 995	89 247	85 228
Säkeitä	2 162 948	1 417 090	893 288
Sanoja	7 346 075	4 259 398	2 599 158
Vuosilta	1644, 1804–1943[2]	1564–1939	1653–1971
Aineiston teksteistä 50 % vuosilta	1889–1923	1884–1915	1930–1947
Maakuntia	13	24	29
Pitäjiä tai kaupunkeja	118	532	622
Tallentajia	4192	1 634	2 413
Pääkielet	pohjois- ja eteläviro, seto	inkeroinen, karjala, lyydi, suomi, vatja	inkeroinen, karjala, suomi
Aineistoa myös kielillä (yksittäisiä tekstejä tai sanoja)	inkerinsuomi, inkeroinen, saksa, suomi, venäjä	ruotsi, latina, venäjä, kreikka	viro, vepsä, vatja, saame, romani, ruotsi
Aineisto	täydentyy	valmis	valmis, saattaa täydentyä
Runotyyppihakemisto	kesken (ei kata koko aineistoa)	kesken (tarve tarkentaa ja tarkistaa)	ei ole (kortisto on)
Runotyyppiotsikoita	2514 (vanhoja 14 818)	7573	-

2 ERAB-aineiston yksi pieni keräelmä on merkitty vahingossa vuodelle 1957 ja korjataan.

Runotyyppiotsikoita, joihin kuuluu vain yksi teksti	828 (vanhoista 10 181)	2827	-
Normalisoitu tietokantaan	runoteksti, paikka, aika, kerääjänimi	paikka, aika, kerääjänimi	paikka, aika, kerääjänimi

Kuva 2. Taulukko ERAB-, SKVR- ja JR-korpusten perusominaisuuksista FILTER-tietokannan pohjalta laskettuna. Kielten esiintymistä ei ole tarkasteltu laskennallisesti, ja tiedot saattavat olla vaillinaisia. Taulukko: Kati Kallio, Maciej Janicki, Jukka Saarinen, Liina Saarlo ja Mari Sarv 2023.

Korpusten järjestelyhistoria

Aineiston pitkien synty- ja järjestelyhistorioiden tuottamat piirteet ja rakenteet ohjaavat helposti tulkintoja sähköisessä muodossakin. Esimerkiksi SKVR-kirjasarjan julkaisujärjestys oli sidoksissa 1900-luvun alussa arvokkaimpana ja tutkimuksen kannalta tarpeellisimpina pidettyihin aineistoihin ja alueisiin (esim. Hautala 1954, 197): julkaisusarja alkoi Vienan Karjalasta, kertovista runoista ja yksittäisten runojen tasolla Samporunostosta, ja viimeisenä ennen vuoden 1997 täydennysosaa tulivat Uudenmaan runot. Osin järjestys määräytyi myös ulkopuolisen rahoituksen mukaan (ks. Krohn 1916). SKVR-korpuksen teksteille annetut sähköiset tunnusnumerot säilyttävät tämän järjestyksen. Jos minkä hyvänsä hakutuloksen järjestää näiden mukaan, määrittyy tarkastelujärjestys SKVR-sarjan julkaisujärjestyksen ja siten myös sen kantaman aikansa arvojärjestyksen mukaisesti. Järjestelyhistoria vaikuttaa sähköisestäkin aineistosta tehtäviin tulkintoihin, joten siitä on tarpeen olla tietoinen.

Suomalaisen Kirjallisuuden Seuran yksi päätarkoitus oli sen perustamisesta vuonna 1831 lähtien edistää kansanrunouden tallentamista. Vuonna 1890 valmistui SKS:n oma talo, jonka jälkeen käsikirjoituksia alkoi kertyä enemmän ja niiden järjestäminen muuttui systemaattisemmaksi. Kokoelmat järjestettiin kerääjittäin ja luokiteltiin karkeasti perinnelajeittain. Kansallisesti keskeisimpinä pidettyjen suullisten kalevalamittaisten runojen julkaisemista yritettiin jo 1800-luvun lopulla (Hautala 1950; *Arkiston avain* 1984). SKS:n kansanrunousarkiston perustamisen jälkeen vuonna 1934 alettiin aineistoja koota sidoksiksi, dokumentoida tarkemmin kerääjäkortistoon ja laatia myös tarkempia perinnelaji- ja paikkakuntakohtaisia kortistoja (Laaksonen ja Saarinen 2004, 7–11; Harvilahti 2013).

Viron Kansanrunousarkisto (Eesti Rahvaluule Arhiiv, ERA) perustettiin vuonna 1927 Viron Kansallismuseon osaksi, kun haluttiin kerätä yhteen ja helpommin saataville jo kertyneet laajat, pääosin yksityiskokoelmissa olleet kansanperinnekokoelmat (ks. Järv 2013). Tämä tarkoitti myös aineistojen kartoittamista, sisältöluetteloiden laatimista sekä runolaulun kopiokortiston ja muiden vastaavien lajikohtaisten kortistojen laatimista tärkeimpinä pidettyjen lajien osalta (ks. Loorits 1932). Aineiston metadataa systematisoitiin ja täydennettiin esimerkiksi tallennuspitäjien osalta ja siihen lisättiin perinnelaji- ja runotyyppimerkinnät. Runolaulukortistosta laadittiin alkuperäiskäsikirjoitusten suojelemisen ja käytön helpottamisen nimissä paikan, keräelmän luovutusajan sekä lajin ja runotyypin mukaiset versiot, joita täydennettiin 1990-luvulle asti. (Sarv ja Oras 2020, 108–109.) Järjestelmä oli hyvin samanlainen kuin SKS:n kansanrunousarkistossa; suomalaisten ja virolaisten kansanrunouden tutkijoiden ja tallentajien välillä oli yhteistyötä Elias Lönnrotista ja Friedrich Reinhold Kreutzwaldista alkaen (Järvinen 2008). Neuvostomiehityksen (1940) seurauksena Viron kansalliset instituutiot lakkautettiin tai pilkottiin ja järjestettiin uudelleen, osa tutkijoista pakeni maasta, ja yhteistyösuhteet ulkomaille katkesivat lähes täysin muutamaksi vuosikymmeneksi. Kansanrunousarkiston toiminta kuitenkin jatkui osana Valtiollista Kirjallisuusmuseota. (Sarv ja Oras 2020; ks. myös

Järvinen 2008; Kansanrunousarkiston ja Kirjallisuusmuseon historiasta ks. Eesti Kirjandusmuuseum n.d.)

SKVR-kirjasarja (1908–1948 ja 1997) laadittiin alkuperäisaineiston säilymisen ja paremman käytettävyyden tueksi: julkaisemisensa jälkeen siitä tuli suomalaisen kalevalamittaisen runouden tutkimuksen keskeisin lähde. Mukaan pyrittiin ottamaan kaikki kalevalamittaiset tai kalevalamittaiseen perinteeseen läheisesti liittyvät suullista alkuperää olevat tekstit. Karsinta tapahtui oman aikansa arvostusten ja aitouskäsitysten perusteella: etenkin kirjallislähtöisiksi tai laulajan itse sepittämäksi epäiltyjä, seksuaalirunoja ja pilkkalauluja sekä käsikirjoituskopioita jätettiin julkaisematta. SKVR:n eri osissa toimitusperiaatteet vaihtelivat jonkin verran. Suurin osa niteistä järjestettiin runotyypeittäin ja pitäjittäin, mutta esimerkiksi Inkerin aineistot niiden variaation vuoksi tallentajittain. Pyrkimyksenä oli runojen julkaiseminen mahdollisimman pitkälti alkuperäislähteitä vastaavassa muodossa. Etenkin runojen tallennus- ja paikkakuntatietoja selvitettiin jälkikäteen eri lähteistä. (SKVR-sarjan esipuheet; Hautala 1950; Mäkelä ja Tarkka 2022.)

SKS:n julkaisemattomien runojen aineistoa alettiin luoda Suomen Akatemian rahoittaman ja Matti Kuusen johtaman hankkeen puitteissa vuosina 1969–1973, jolloin poimittiin Kansanrunousarkiston käsikirjoitusaineistoista täydentävää listausta julkaisematta jääneistä kalevalamittaisista runoista, joita sitten 1990-luvun alkupuolelle asti kopioitiin SKS:n kansanrunousarkiston konekirjoittajien voimin kahdeksi kortistoksi. (Keskustelut Senni Timosen kanssa, SKSÄ 2023:3; SKSÄ 2023:30.) Mukaan otettiin SKVR-sarjasta pois jätettyjen runojen ohella myös myöhemmin kertynyttä kalevalamittaista aineistoa sekä muita lajeja, kuten ”arvoitukset, uusimittaiset laulut, pyrkimäluvut, itkuvirret, luonnonäänien jäljittelyt ja joiut” siltä osin, kun näitä ei oltu kopioitu vielä omiin kortistoihinsa (*Arkiston avain* 1984, 26).

Jukka Saarinen, Senni Timonen, Sakari Korpikallio ja Kati Kallio ovat vertailleet alkuperäistä ATK-poimintalistaa ja poimintahankkeen muistiinpanoja julkaisemattomien runojen kortistosta tehtyyn sähköiseen JR-korpukseen. Näyttää siltä, että poiminta kohdistui joihinkin keräelmiin melko satunnaisesti ja jäi ilmeisesti kesken. Lisäksi sellaiset aineistot, joille ei aikanaan ole annettu kalevalamittaiseen perinteeseen viittaavaa perinnelajikoodia – kuten osa rahvaanrunoista, suoraan epäaitoon f-aineistoon ilman perinnealajikoodia luokitellut runot tai kirjallisuusarkiston kokoelmiin erotetut aineistot – eivät osuneet läpikäytävään aineistoon ollenkaan. Senni Timosen (SKSÄ 2023:3; SKSÄ 2023:30) arvion mukaan kopiointivaiheessa konekirjoittajille ei ollut aina kirjallisia ohjeita, mikä selittänee esimerkiksi aineiston osin virheellisiä paikkakuntamerkintöjä sekä typografiaan ja kontekstitietojen määrään liittyviä vaihtelevia ratkaisuja. JR-korpus on siten laaja, SKVR-korpusta monin tavoin täydentävä aineisto, mutta ei kattava, yhtenäinen tai tarkistettu.

Virossa runojen pitäjä kerrallaan etenevä julkaiseminen *Vana Kannel* -sarjassa (’Vanha kantele’) alkoi jo vuonna 1876, oli pysähdyksissä vuodet 1941–1985 ja jatkuu yhä. Julkaisu vastaa monilta piirteiltään SKVR-sarjaa, mutta siinä julkaistaan sekä runotekstejä että nuotteja. (Ks. Saarlo 2012; Järv 2016, 33.) Myös muita runo- ja sävelmäjulkaisuja on tehty: tutkimuksen, kortistojen järjestämisen ja runotyyppien hahmottamisen kannalta erityisen vaikutuksellinen on ollut Ülo Tedren antologia *Eesti Rahvalaulud* (1969–1974; tästä ja muista virolaisten runojen julkaisuista ks. myös Kuusi 1977; Saarlo 2012).

Virossa arkistoaineistoja alettiin jo 1990-luvun puolivälissä digitoida ja muodostaa pääosin lajikohtaisiksi tietokannoiksi, jotka alkoivat korvata manuaalisia kortistoja (Järv ja Sarv 2014; Järv 2016; Sarv ja Oras 2020, 110; SKS:n digitointihankkeista ks. Klemettinen 2006). Vuonna 1998 SKS:n kansanrunousarkisto ehdotti Viron kansanrunousarkistolle yhteistyötä runoaineistojen digitoimiseksi (Saarinen 2001; 2006). Jukka Saarisen ja Senni Timosen (SKSÄ 2023:3; SKSÄ 2023:30) tietojen mukaan aloite tuli alkujaan virolaiselta akateemikko Arvo Krikmannilta, ja erityisesti SKS:n pääsihteeri Urpo Vento piti yhteistyötä tärkeänä. Matti Kuusi ja Ülo Tedre (1979) olivat jo aikaisemmin pohtineet virolaisia ja suomalaisia runoaineistoja vertailevaa projektia, jota varten tehtiin 1980-luvun alkupuolella Kuusen SKVR-hakemistoa käsitelleen Suomen Akatemian hankkeen yhteydessä myös pohjatyötä. Siinä missä Kuusi keskittyi Timosen mukaan julkaisuhankkeisiin, Vento katsoi tärkeämmäksi saattaa aineistoja sähköiseen muotoon.

Digitointi oli edullisempaa tehdä Virossa, ja samalla Viron Kansanrunousarkistoon saatiin hankittua siellä tarvittuja laitteita. Työ aloitettiin 2000-luvun alussa SKVR-aineistosta, jonka Arvo Krikmannin johtama ryhmä kuvasi suoraan kirjasarjasta ja muutti tekstimuotoon. Sama ryhmä digitoi sitten samoja periaatteita käyttäen tekstimuotoon myös SKS:ssa kuvatun julkaisemattomien runojen kortiston sekä konekirjoitetussa muodossa olevan Viron kansanrunousarkiston runolaulukortiston. (Saarinen 2006; Sarv ja Oras 2020; SKSÄ 2023:3; SKSÄ 2023:30.) Liina Saarlon laskelman mukaan virolaisesta runoaineistosta oli digitointia aloitettaessa julkaistu alle kymmenen prosenttia, josta noin viisi prosenttia *Vana Kannel* -sarjassa. Virolaisten runojen digitointi oli siis mielekkäintä tehdä kortiston pohjalta.

Arvo Krikmann rakenteisti tekstimuotoon digitoidun SKVR-aineiston xml-muotoon keskusteltuaan rakenteesta Jukka Saarisen ja Pasi Klemettisen kanssa. Jukka Saarinen tarkisti ja korjasi aineistoa, antoi kullekin tekstille ID-tunnisteen, normaalisti kerääjä-, paikka- ja vuositiedot yksiselitteisiksi attribuuteiksi ja yhtenäisti erikoismerkkien koodiston Unicode-pohjaiseksi. (Saarinen 2006; ks. myös Sarv ja Oras 2020.) Julkaisemattomien runojen (JR) osalta metatietojen ja erikoismerkkien normaalistamista ei tehty, ja korpus oli vuoteen 2022 asti käytettävissä vain SKS:n arkiston tekstitiedostoina sekä manuaalisena kortistona. Sen metatietokenttien merkinnät vaihtelevat: yksi paikkakunta voi esimerkiksi esiintyä useilla nimivarianteilla. FILTER-hankkeessa Jukka Saarinen ja Maciej Janicki ovat normaalistaneet JR-korpuksen paikka-, aika ja kerääjänimitiedot, ja Saarinen on tehnyt aineistoon jokin verran korjauksia. Korpusta ei ole kuitenkaan kattavasti tarkistettu, ja etenkin paikkatiedoissa on yhä epäjohdonmukaisuuksia.

Virolaista runokortistoa alettiin digitoimisen jälkeen toimittaa keräelmä kerrallaan ERAB-korpukseksi arkiston tutkijoiden voimin, tekijöinä etenkin Janika Oras, Liina Saarlo ja Mari Sarv (Sarv ja Oras 2020, 110–111). Virossa sähköistä korpusta ei siis alettu tehdä minkään julkaisusarjan pohjalta, vaan keräelmittäin yleensä arkiston vanhimmista aineistoista lähtien, kortistojen kautta edeten ja – toisin kuin SKS:n julkaisemattomien runojen kohdalla – kaikki tekstit käsikirjoituksista tarkistamalla. Samalla on yhtenäistetty metatietoja ja tekstien kirjoitusasua, murremuodot säilyttäen. Käytössä on sekä alkuperäinen että normaalistettu tekstimuoto. Kirjoitusasun yhtenäistäminen tekee virolaisesta korpuksesta monikäyttöisemmän ja tekstihakujen kannalta helpomman. Paikka- ja kerääjätiedot on normalisoitu. Suomalaisesta aineistosta poiketen ERAB-korpuksen metadatassa on myös standardoidut kentät esimerkiksi tiedoille laulajasta ja kerääjän tai laulajan runolle antamasta otsikosta, ja runoteksteissä refrengit ja kommentit on merkitty erikseen. Vuodesta 2016 lähtien ERAB-tietokannassa on

ollut myös mahdollista sijoittaa hakutulokset kartalle. (Sarv ja Oras 2020, 111–112.) Tällä hetkellä korpuksessa on noin kaksi kolmannesta arkiston runolauluteksteistä ja lisäksi noin 6000 välimuotoista tai riimillistä laulua sekä joitain proosatekstejä. (*Regilaulude andmebaas*: andmebaasist.) Toisin kuin SKVR-korpuksessa, mukaan ei ole otettu kauempana kalevalamitasta olevia loitsutekstejä, vaan nämä ovat arkistossa omana kokonaisuutenaan, joka on puolestaan digitoitu Mare Kõivan (ks. 2019) hankkeissa.

Syntyhistorialtaan ja luonteeltaan ERAB-, SKVR- ja JR-korpukset eroavat siis toisistaan, vaikka aineistojen synty, järjestäminen ja digitoiminen on tapahtunut keskinäisessä vuorovaikutuksessa. Keskinäisistä eroistaan huolimatta korpukset on alusta lähtien rakennettu sopimaan yhteen toistensa kanssa. (Harvilahti 2013; Sarv ja Oras 2020, 110–111.)

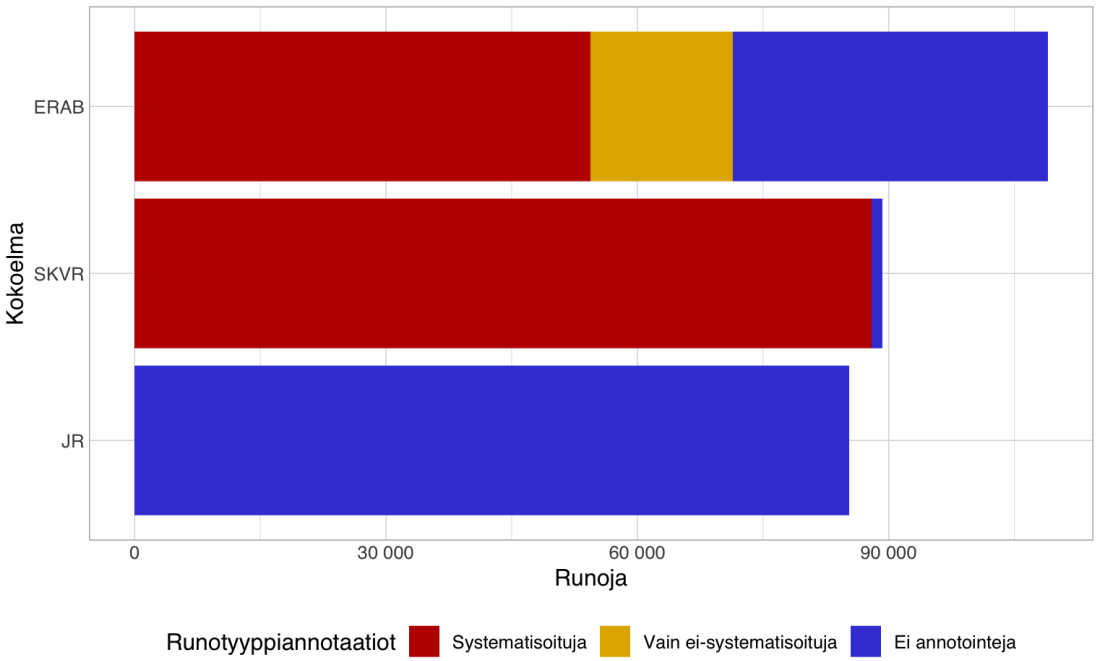
Kolmen korpuksen perusrakenne on yksikkötasolla sama: metatiedot (paikka, aika, kerääjänimi tai aineiston luovuttanut toimija, arkistosignum), vapaa tekstikenttä ja itse runoteksti säkeittäin aseteltuna sekä joskus vapaata tekstiä, alaviitteitä tai kommentaaria myös runotekstin jälkeen. ERAB-korpuksessa metatietokenttiä on enemmän kuin kahdessa muussa. Runoteksti noudattaa korpuksissa yleensä alkuperäisen lähteen kieliasua – SKVR-korpuksessa toimittajien täydennykset on merkitty hakasulkeisiin ja ERAB-korpus sisältää myös normalisoidun tekstiversion. SKVR-tietokannassa ja FILTER-hankkeessa on käytössä Jukka Saarisen laatima normaalistussäännöstö, joka tekstihakujen helpottamiseksi poistaa diakriittiset merkit ja erikoismerkit (ks. *SKVR-tietokanta:* Ohjeita).

Sekä ERAB-korpuksessa että SKVR-korpuksessa on käytössä runotyyppihakemisto (ks. Sarv ja Oras 2020; *SKVR-tietokanta*: SKVR:n runohakemisto). SKVR-hakemiston pohjana ovat SKVR-kirjasarjan nidekohtaiset hakemistot, mutta etenkin lyriikan, loitsujen ja häälaulujen osalta aineisto on nykyistä hakemistoa varten analysoitu uusiksi. Työ aloitettiin 1980-luvun alussa Matti Kuusen akatemiahankkeessa ja sitä jatkettiin SKS:n kansanrunousarkiston puitteissa. JR-korpuksessa hakemistoa ei ole – arkistossa on kuitenkin aineistosta paikkakuntakortiston jälkeen omana projektinaan aloitettu, SKVR-niteiden hakemistoja mukaileva runotyyppikohtainen kortisto. (SKSÄ 2023:3; SKSÄ 2023:30.) ERAB-korpukseen luodaan paraikaa systematisoitua luokitusta, joka valmiilta osaltaan on jo käytössä. (Sarv ja Oras 2020, 112.) Työn pohjana on Ülo Tedren antologiassa *Eesti Rahvalaulud* (1969–1974) käyttämä jäsennys (ks. https://www.folklore.ee/laulud/erla/indeks1.html). Lisäksi suuressa osassa aineistoa on arkistokortistosta mukaan otettuja systematisoimattomia, vaihtelevampia tyyppinimiä.

Suomalainen ja virolainen runohakemisto on tehty erilaisten luokitushistorioiden ja -periaatteiden pohjalta. Esimerkiksi virolaisessa hakemistossa *Loomine* (’luominen’) löytyy lyyrisen epiikan – myyttisten laulujen alta, suomalaisessa taas *Maailmansynty* kertovien runojen – epiikan alta. Osin luokittelukäytännöt vaihtelevat myös lajeittain (ks. *SKVR-tietokanta*: SKVR:n runohakemisto).

Tyyppiluokitetun eli annotoidun aineiston määrä eri korpuksissa näkyy kuvassa 3 (seuraavalla sivulla), jossa luokittelematon aineisto on merkitty sinisellä, nykyisiin tyyppihakemistoihin kuuluva osuus punaisella ja ainoastaan vanhat systematisoimattomat tyyppinimet sisältävä aineisto keltaisella.

ERAB-korpus hakemistoineen täydentyy jatkuvasti ja julkaistuja osia voidaan myös korjata. Sähköinen SKVR-korpus pidetään mahdollisimman lähellä alkuperäistä kirjasarjaa: ainoastaan ilmeiset virheet korjataan. Korpuksen runotyyppihakemistoa kuitenkin voidaan



Kuva 3. Runotyyppihakemistot ja niiden kattavuus ERAB-, SKVR- ja JR-korpuksissa. Kuva: Eetu Mäkelä 2023.

täydentää ja korjata. Julkaisemattoman runoaineiston (JR) osalta muutoksia tehdään tarvittaessa, ja aineistoa voidaan lisätä korpukseen. SKVR-data on avoimesti saatavilla SKS:n organisaatiotilillä GitHubissa, ja toiveena on saada myös ERAB ja JR avoimiksi korpuksiksi.

Aineiston käsittely FILTER-hankkeessa

Maciej Janicki ja Eetu Mäkelä ovat vuosina 2020–2022 muiden FILTER-hankkeen jäsenten (Saarlo, Sarv, Saarinen, Kallio) kanssa keskustellen koonneet SKVR-, ERAB- ja JR-korpukset yhdeksi tietokannaksi. Tietokantamuoto mahdollistaa aineiston käsittelemisen ja vertailemisen laskennallisesti. Tietokantaan on myös lisätty *Kalevala* (1835 ja 1849), *Kanteletar*, *Kalevipoeg* ja *Ilo-Laulu Jeesuksesta*, joita ei kuitenkaan käsitellä tässä artikkelissa. Aineistojen käyttöönotto on edellyttänyt jonkin verran erilaisten kenttien, merkintätapojen ja kuratointihistorioiden selvitystyötä.

Käytännössä hankkeen työ on edennyt aineistoa, sen nykytutkimusta ja oppihistoriaa tuntevien folkloristien ja laskennallisia menetelmiä tuntevien tietojenkäsittelytieteilijöiden välisenä vuoropuheluna. Lähtökohtana on pyrkiä lomittamaan aineiston määrällistä ja laadullista tarkastelua: määrällisiäkin tuloksia on voitava tarkastella, testata ja arvioida myös yksittäisten tekstien tasolla. Näissä auttavat Eetu Mäkelän luoma, muissakin hankkeissa käytetty hakuliittymä *Octavo* sekä Maciej Janickin luoma *Runoregi*-liittymä, joka antaa lähiluettavaksi runojen samankaltaisuuslaskentojen tuloksia (näistä tarkemmin ks. Kallio ja Mäkelä 2019; Kallio ym. 2020; Janicki 2022; Kallio ym. 2022; Janicki ym. 2023). Nimi *Runoregi* on yhdistelmä vanhaa kansanrunoa kuvaavista sanoista *runolaulu* (suomi, karjala) ja *regilaul* (viro), ja se viittaa myös karjalaisessa ja inkeriläisessä perinteessä käytettyyn lauluja kuljettavan *runoreen* käsitteeseen (karjalan kielessä usein *regi*). Octavoa ja Runorekeä on käytetty myös tämän artikkelin määrällisiä tuloksia pohdittaessa ja tarkistettaessa.

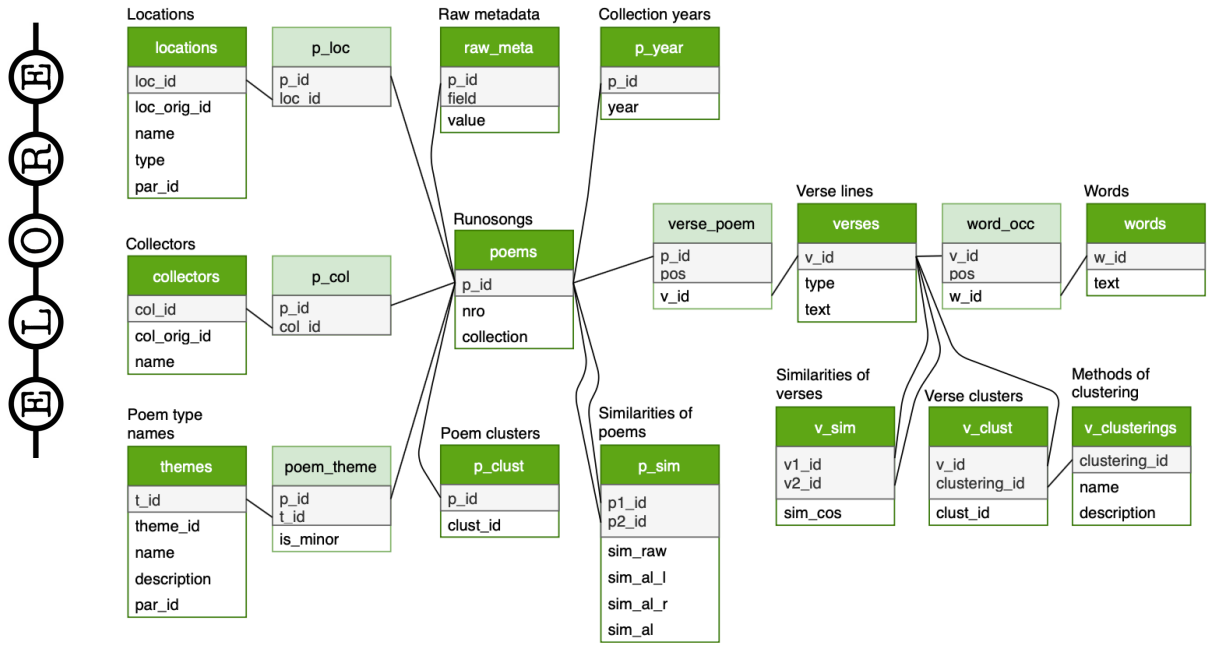
Välineiden tasolla tavoitteena ei ole luoda yhtä suurta analyysialustaa, vaan ottaa tarpeen mukaan käyttöön erilaisia välineitä, joita voi kytkeä toisiinsa. Octavo-hakuliittymä antaa haluttaessa runokohtaiset linkit runosäkeiden ja tekstien samankaltaisuutta tunnistavaan Runoregi-liittymään, jossa on myös mahdollista selata tarkemmin SKVR- ja ERAB-runotyyppihakemistoja. Kartta- ja lajijakaumakuvia antavassa visualisaatiosovelluksessa (FILTER Visualizations) on mahdollisuus tehdä esimerkiksi Octavon hakusyntaksia tai FILTER-tietokannan SQL-syntaksia noudattavia kyselyitä, ja siihen on pääsy myös suoraan joistain Runoreen näkymistä.

Perinteisen SKVR- tai arkistoviitteen sijaan aineiston käsittelyssä on olennaisin arkistojen jokaiselle runotekstille antama yksilöllinen ID- eli tunnusnumero. ERAB-korpuksessa tunnusnumerot pohjautuvat keräelmäkohtaisiin arkistoviitteisiin, joten niitä on eri muotoisia. SKVR-korpuksen kohdalla tunnusnumerot perustuvat perinteiseen viitteeseen. Esimerkiksi skvr05308170 tarkoittaa SKVR-sarjan (skvr) osan V (Pohjois-Inkeri, 05) niteen 3 (3) runoa 817 (08170).[3] Tunnusnumeron avulla saa etsittyä helposti runon sivukohtaisen osoitteen myös alkuperäisessä SKVR-tietokannan osoiterivin kautta (esim. https://skvr.fi/poem/skvr05308170), ja esimerkiksi Runoreessä niitä voi käyttää tiettyjen näkymien muokkaamiseen selaimen osoiterivillä. JR-korpuksen osalta tunnusnumerot ovat juoksevia numeroita, joissa on joitain poikkeamia suhteessa aineiston järjestykseen arkiston pääkortistossa.

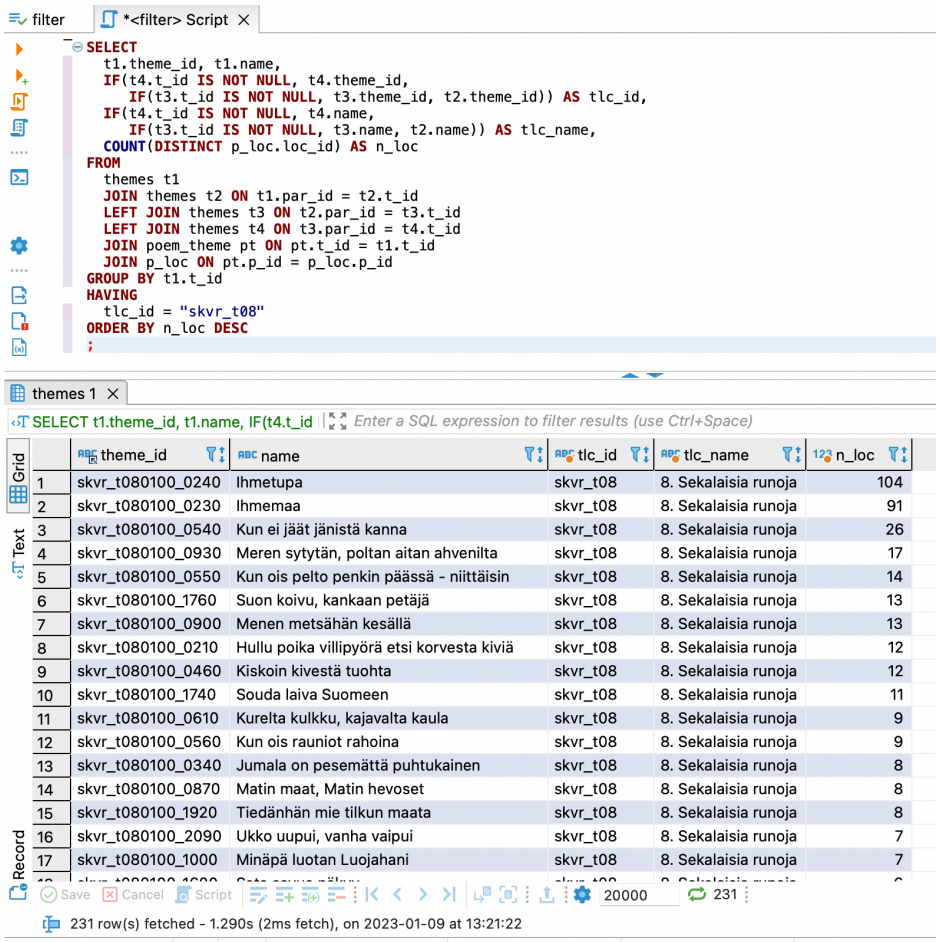
Hankkeen karttapohjan on tehnyt Samppa Mäkelä vuosina 2020 ja 2023 muokkaamalla Kotuksen sähköistä pitäjäkohtaista aineistoa SKS:n perinnealuejakoa (ks. Nissilä ym. 1970; Sarmela 2007, 591–592) ja ERAB-tietokannan pitäjäjakoa vastaavaksi. Pitäjät vastaavat 1900-luvun alkupuoliskon tilannetta hiemain alueittain vaihdellen. Vanhemmat tai uudemmat tekstit eivät aina istu karttapohjaan virheettä: yksikin pitäjä on voinut aikojen kuluessa vaihdella kooltaan ja nimeltään, ja aineiston pitkään tallennusajanjaksoon sisältyy erilaisia pitäjien jakoja ja yhdistämisiä. Kartta on geokoodattu eli sidottu pohjakartan koordinaatistoon, mutta pitäjärajat eivät ole kovin tarkkoja. Vermlannin ja Tverin alueet sekä Baltian kielisaarekkeet ovat vielä piirtämättä ja etenkin Ruijan alue korjaamatta.

Hankkeessa käytetään yhdeksi FILTER-tietokannaksi koostettua kokonaisaineistoa, joka on muodoltaan csv-tiedostoista ladattava relaatiotietokanta (MariaDB). Siinä missä csv-muotoinen tieto on helposti siirrettävissä käyttöympäristöstä toiseen, mahdollistaa tietokanta aineiston tarkastelemisen monimutkaistenkin kyselyiden (*query*; ks. Kuva 5 sivulla 69) avulla. Relaatiotietokannassa tietoa käsitellään toisiinsa matemaattisissa suhteissa olevina kokonaisuuksina. Käytännössä tietokanta koostuu toisiinsa kytketyistä taulukoista (*table*; ks. Kuva 4 seuraavalla sivulla). Taulukoiden rivit vastaavat aineiston yksiköitä (*entry*) – taulukosta riippuen vaikkapa runoa, säettä, sanaa, kerääjää tai paikkakuntaa – ja jokainen sarake antaa tietynlaisen tiedon tästä yksiköstä. Tällaisille yksiköille on myös annettu omat numeromuotoiset ID-tunnuksensa. Kyselyissä käytetään hyväksi näitä tietoja ja taulukoiden suhteita toisiinsa. Tietokantaa on helppo täydentää vaikkapa uusia laskentatuloksia sisältävillä taulukoilla. Esimerkiksi Runoregi-käyttöliittymän pohjana olevat samankaltaisuuslaskelmat on liitetty tietokantaan, joten niitä on mahdollista käyttää hyväksi myös tietokantapohjaisissa kyselyissä. Tietokantaa voi käyttää erilaisten käyttöliittymien kautta, esimerkiksi avoimen lähdekoodin ilmaisia hallintatyökaluja (kuten HeidiSQL tai DBeaver) tai R-ohjelmointiympäristöä käyttäen.

3 SKVR-tunnusnumeron kolmas numero viittaa kyseisen SKVR-osan niteeseen tai juoksevaan numerosarjaan, joita voi yhdessä niteessä olla useampia: esimerkiksi loitsut ovat kirjasarjojen niteissä usein omana numerosarjanaan. Oma numeronsa on myös annettu niille runoille, joilla ei ollut kirjasarjassa numeroa ollenkaan.



Kuva 4. FILTER-tietokannan taulukkorakenne keskeisimpien taulukoiden osalta. Kuva: Susanna Mett 2023.



```
SELECT
  t1.theme_id, t1.name,
  IF(t4.t_id IS NOT NULL, t4.theme_id,
     IF(t3.t_id IS NOT NULL, t3.theme_id, t2.theme_id)) AS tlc_id,
  IF(t4.t_id IS NOT NULL, t4.name,
     IF(t3.t_id IS NOT NULL, t3.name, t2.name)) AS tlc_name,
  COUNT(DISTINCT p_loc.loc_id) AS n_loc
FROM
  themes t1
  JOIN themes t2 ON t1.par_id = t2.t_id
  LEFT JOIN themes t3 ON t2.par_id = t3.t_id
  LEFT JOIN themes t4 ON t3.par_id = t4.t_id
  JOIN poem_theme pt ON pt.t_id = t1.t_id
  JOIN p_loc ON pt.p_id = p_loc.p_id
GROUP BY t1.t_id
HAVING
  tlc_id = "skvr_t08"
ORDER BY n_loc DESC
;
```

themes 1

SELECT t1.theme_id, t1.name, IF(t4.t_id — Enter a SQL expression to filter results (use Ctrl+Space)

	theme_id	name	tlc_id	tlc_name	n_loc
1	skvr_t080100_0240	Ihmetupa	skvr_t08	8. Sekalaisia runoja	104
2	skvr_t080100_0230	Ihmemaa	skvr_t08	8. Sekalaisia runoja	91
3	skvr_t080100_0540	Kun ei jäät jänistä kanna	skvr_t08	8. Sekalaisia runoja	26
4	skvr_t080100_0930	Meren sytytän, poltan aitan ahvenilta	skvr_t08	8. Sekalaisia runoja	17
5	skvr_t080100_0550	Kun ois pelto penkin päässä - niittäisin	skvr_t08	8. Sekalaisia runoja	14
6	skvr_t080100_1760	Suon koivu, kankaan petäjä	skvr_t08	8. Sekalaisia runoja	13
7	skvr_t080100_0900	Menen metsähän kesällä	skvr_t08	8. Sekalaisia runoja	13
8	skvr_t080100_0210	Hullu poika villipyörä etsi korvesta kiviä	skvr_t08	8. Sekalaisia runoja	12
9	skvr_t080100_0460	Kiskoin kivestä tuohta	skvr_t08	8. Sekalaisia runoja	12
10	skvr_t080100_1740	Souda laiva Suomeen	skvr_t08	8. Sekalaisia runoja	11
11	skvr_t080100_0610	Kurelta kulkku, kajavalta kaula	skvr_t08	8. Sekalaisia runoja	9
12	skvr_t080100_0560	Kun ois rauniot rahoina	skvr_t08	8. Sekalaisia runoja	9
13	skvr_t080100_0340	Jumala on pesemättä puhtukainen	skvr_t08	8. Sekalaisia runoja	8
14	skvr_t080100_0870	Matin maat, Matin hevoset	skvr_t08	8. Sekalaisia runoja	8
15	skvr_t080100_1920	Tiedänhän mie tilkun maata	skvr_t08	8. Sekalaisia runoja	8
16	skvr_t080100_2090	Ukko uupui, vanha vaipui	skvr_t08	8. Sekalaisia runoja	7
17	skvr_t080100_1000	Minäpä luotan Luojahani	skvr_t08	8. Sekalaisia runoja	7

Save Cancel Script 20000 231

231 row(s) fetched - 1.290s (2ms fetch), on 2023-01-09 at 13:21:22

Kuva 5. Tietokantakysely ja vastaustaulukon alku siitä, kuinka monen pitäjän alueella SKVR-aineiston hakemistokategoriaan *Sekalaisia runoja* kuuluvat runotyypit esiintyvät, käyttöliittymänä DBeaver. Kysely: Maciej Janicki, kuva: Kati Kallio 2023.

Aineiston käsittely tietokannaksi on systematisoitu ja dokumentoitu: jos aineistoon tulee päivityksiä, uusi versio voidaan ajaa FILTER-tietokantaan helposti. Prosessi koostuu tällä hetkellä kahdesta osasta. 1) Formaattien konversio muuttaa erilaisista lähteistä tulevan datan (SKVR ja JR: XML-muoto; ERAB: tietokantavedos; kirjalliset runot: tekstitiedostot Project Gutenbergista) joukoksi rakenteeltaan standardisoituja tietokantatauluja ja kytkee eri korpusten taulut toisiinsa. Tämä prosessi kestää tavallisellakin tietokoneella vain muutaman minuutin. 2) FILTER-tietokantaan kuuluu myös samankaltaisuuslaskentojen tuloksia, jotka päivitetään aina aineistopäivitysten jälkeen. Koko FILTER-aineiston säetason samankaltaisuuden laskeminen ja klusterointi tämän pohjalta kestää joitain tunteja, runotason samankaltaisuuksien laskeminen (*alignment-based poem similarity*) taas *Tieteen tietotekniikan keskus* CSC:n tarjoaman GPU klusterin (https://docs.csc.fi/computing/systems-puhti/) kauttakin noin 70 tuntia, tavallisella tietokoneella kauemmin. (Laskentamalleista keskeisempiä on kuvattu artikkeleissa Janicki ym. 2023 ja Janicki 2022; ks. myös https://github.com/maciejjan/matrix-align/.)

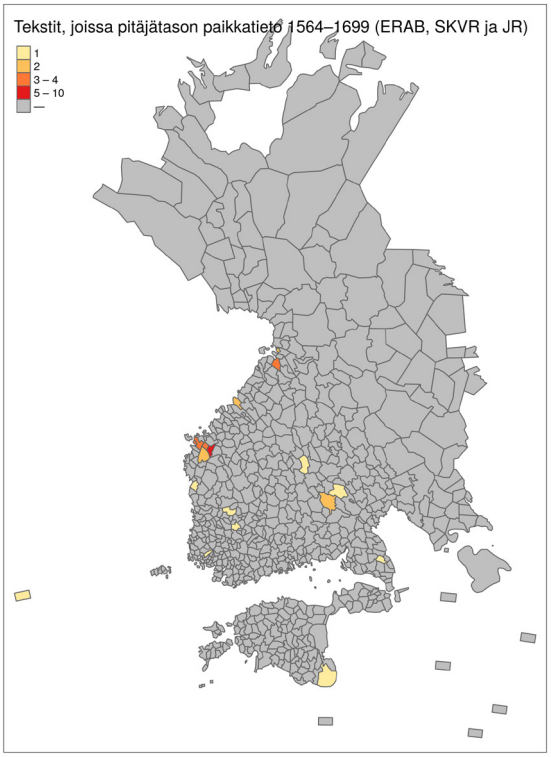
Tietokannan rakennetta ja käyttömahdollisuuksia kuvataan tarkemmin tekeillä olevassa FILTER-dataoppaassa, johon tullaan liittämään myös itse aineistoa tarkemmin kuvaavat osiot (Janicki (tekeillä)). Aineiston ja koodien parissa tehtävä työ dokumentoidaan tarkemmin GitHub-alustalle, joka pitää tallessa myös versiohistorian. Viimeistään FILTER-hankkeen päättyessä aineiston käsittelyssä käytetyt koodit ovat julkisia ja rajoituksetta muidenkin käytettävissä.

Tallennushistoria määrällisestä näkökulmasta

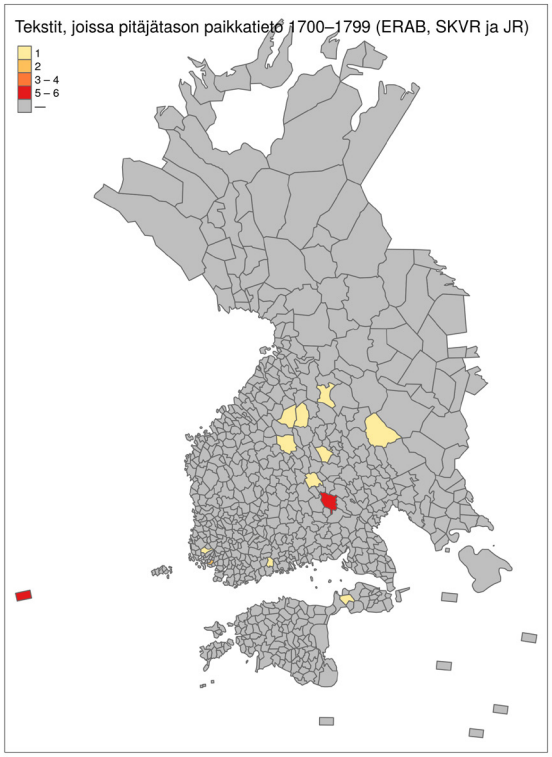
Kalevalamittaisen runouden eli runolaulun (viroksi *regilaul*) aineistoa on tallentunut 1500-luvun lopulta lähtien. Eri aikakausia ovat luonnehtineet erilaiset suulliseen perinteeseen ja sen dokumentoimiseen ja käyttöön kohdistuneet, esimerkiksi kansallisuusaatteisiin, politiikkaan ja modernisaatioon kytköksissä olevat intressit, mutta yhtä lailla myös ajan myötä muuttuvat tieteelliset tulkintakehykset ja kiinnostuksen kohteet. (Tallennus- ja oppihistoriasta ja näiden taustoista esim. Sarajas 1956; Wilson 1976; Tarkka 1989; Jaago 1999; Sihvo 2003; Anttonen 2005; Kuutma ja Jaago 2005; Kuutma 2006; Jaago 2010; Valk 2010; Valk 2014; Kalkun 2015; Mikkola 2021.) Tallentamisen käytännöt ja tavoitteet sekä Suomessa että Virossa kytkeytyvätkin paitsi kansallisvaltioiden rakentamisen transnationaaliin historiaan, myös folkloristiikan kansallisiin kehityskulkuihin ja kansainväliseen oppihistoriaan, joita on viime vuosikymmeninä tarkasteltu monesta eri näkökulmasta (esim. Bendix 1997; Bauman ja Briggs 2003; Kurki 2004; Bendix ja Hasan-Rokem 2012; Kuutma 2015; Tarkka, Haapoja-Mäkelä ja Stepanova 2019; aineiston perusselvityksistä esim. Krohn 1916; Väisänen 1917; Haavio 1931; Hautala 1950; Laugaste 1975). Nämä tekijät ovat vaikuttaneet myös siihen, miten ja minkälaista aineistoa eri aikoina ja eri alueilta on tallentunut. Toisaalta tallentumiseen on vaikuttanut myös se, minkälaisia paikalliset runokulttuurit ovat eri aikoina olleet (esim. Hakamies 1990; Siikala 2000; Timonen 2004; Tarkka 2005; Oras 2008).

Tallennusten määrällisesti runsaimmat jaksot osuvat kalevalamittaisen runouden aineistoissa osin eri kohdille ja painottuvat maantieteellisesti eri seuduille (Kuvat 6–10).

Ruotsin antikviteettikollegio kehotti muinaismuistojen tallentamiseen jo 1600-luvulta alkaen (Sarajas 1956). SKVR-aineiston vanhin teksti on tilikirjaan tallentunut loitsu vuodelta 1564, 1600-luvulta puolestaan on JR- ja SKVR-korpuksessa 49 tekstiä, pääosin oikeustapauksiin liittyneitä loitsuja – osa samojen tapausten eri reittejä kokoelmaan tulleita kopioita – sekä



Kuva 6. Vuosina 1564–1699 ERAB-, SKVR- ja JR-korpuksiin tallentunut aineisto, jossa on pitäjätason paikkatieto. Aikajaksolta 11 tekstiä on ilman pitäjätietoa. Aineisto sisältää joitain kaksoiskappaleita. Kuva: FILTER / Maciej Janicki ja Kati Kallio 2023.

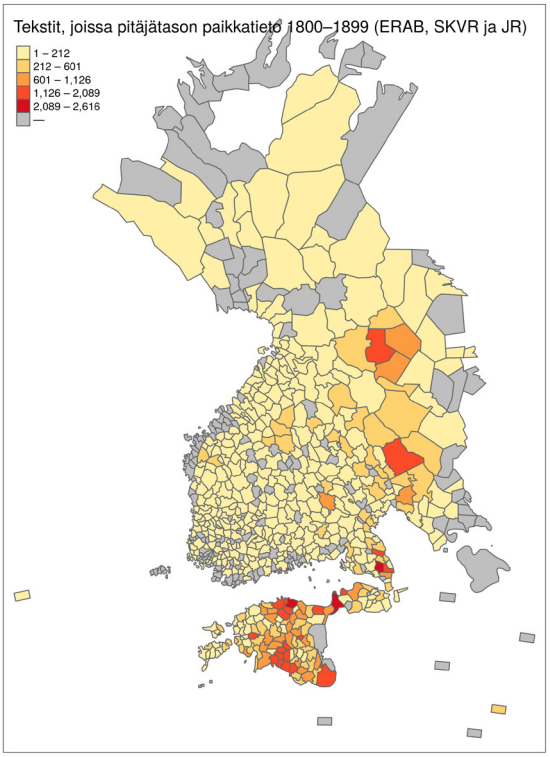


Kuva 7. 1700-luvulla ERAB-, SKVR- ja JR-korpuksiin tallentunut aineisto, jossa on pitäjätason paikkatieto. Aikajaksolta 747 tekstiä on ilman pitäjätietoa. Kuva: FILTER / Maciej Janicki ja Kati Kallio 2023.

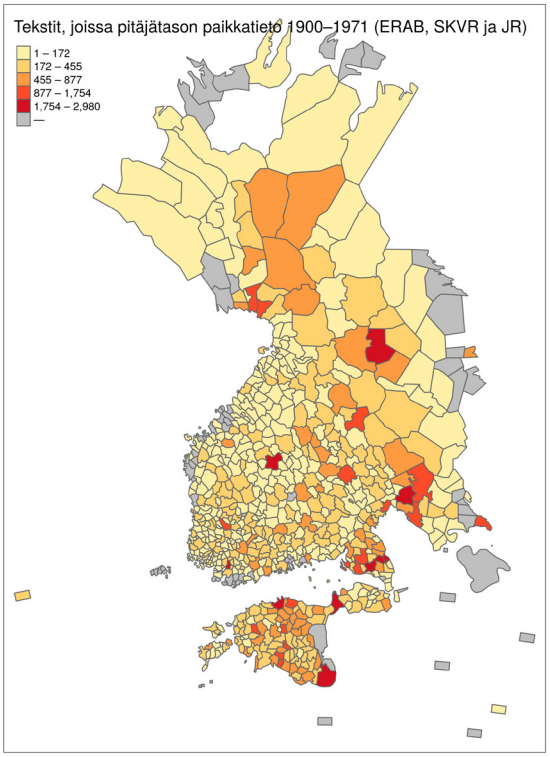
muutama Piispa Henrikin surmavirteen liittyvä tieto. Suomen, mutta etenkään Viron osalta kaikki vanhimmat tunnetut aineistot eivät sisälly sähköisiin korpuksiin (näistä ks. Laugaste 1963; 1975). 1800-lukua edeltävältä ajalta on ERAB-korpuksessa vain yksi Viktor Steinin käsikirjoituskopiona Jakob Hurtin kokoelmaan 1870-luvulla päätynyt, alkujaan vuonna 1644 julkaistu kylvöloitsu Võrumaalta. (Kuva 6.)

1700-luvulta alkaen Turun yliopiston piirissä toimineet suomenkieliset oppineet tallensivat suullisia runoja pääasiassa luterilaisen, Ruotsiin kuuluneen Suomen alueelta (Sarajas 1956). 1700-luvulle on JR-korpuksessa merkitty yksitoista rahvaanrunoa tai oppitekoiselta vaikuttavaa runoa ja SKVR-korpuksessa 754 runoa, joista suurin osa loitsuja (551 tekstiä), mutta mukana on myös epiikkaa, lyriikkaa ja sananlaskuja. Kuusi tekstiä on C. A. Gottlundin Ruotsin suomalaisalueilta saamia käsikirjoituksia, yksi taas vatjalainen teksti 1700-luvun lopulla julkaistusta Liivinmaan kuvauksesta. Aikakauden tallenteista 377 tekstistä ei tiedetä tallennuspaikkaa, ja 364 tekstiä on paikannettu ainoastaan maakuntatasolle. Tämän vuosisadan aineistosta kartta ei siten anna kattavaa kuvaa.

Suomalaiset kokoelmat alkoivat karttua 1800-luvun alkupuoliskolla vauhdikkaammin etenkin Itä-Suomesta ja Venäjän Karjalasta. Lönnrotin Kalevalan ensimmäisen ja toisen version taustalla oli laaja tallennustyö etenkin näillä alueilla, ja Kalevalaa pidettiinkin jonkin aikaa edustavana otoksena kansanrunoudesta. Lisää runoja tallennettiin ensin Kalevalan toista painosta varten, sitten teoksen taustalla olevan laajan kansanrunouspohjan dokumentoimiseksi. (Ks. esim. Hämäläinen 2012; Tarkka, Haapoja-Mäkelä ja Stepanova 2019.) Vuosisadan



Kuva 8. 1800-luvulla ERAB-, SKVR- ja JR-korpuksiin tallentunut aineisto, jossa on pitäjätason paikkatieto. Aikajaksolta 3216 tekstiä on ilman pitäjätietoa. Kuva: FILTER / Maciej Janicki ja Kati Kallio 2023.

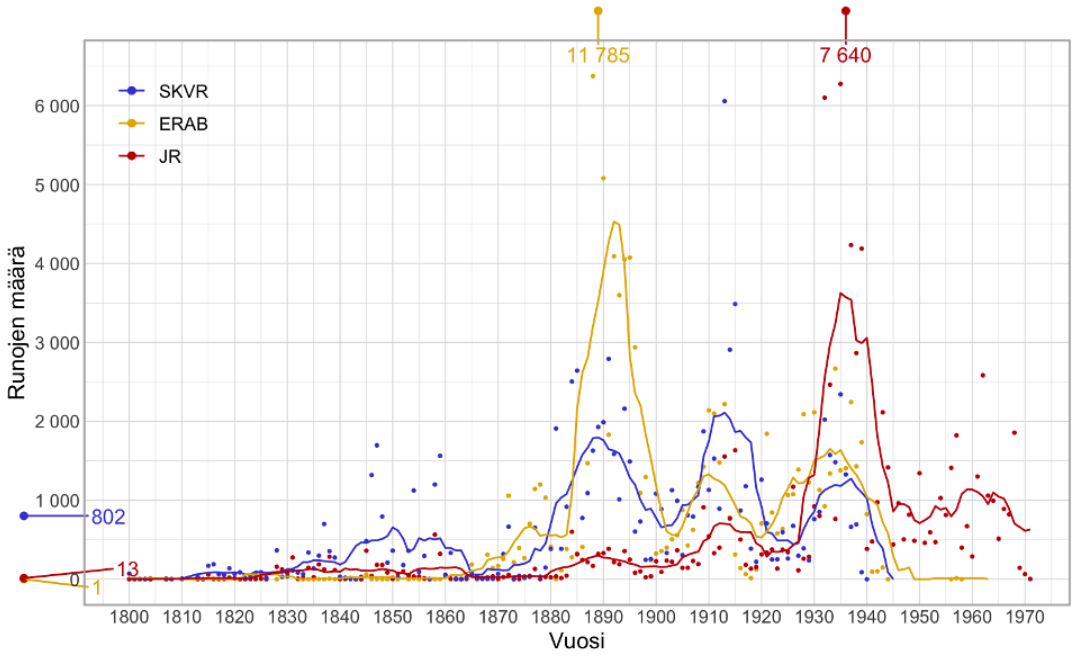


Kuva 9. 1900-luvulla ERAB-, SKVR- ja JR-korpuksiin tallentunut aineisto, jossa on pitäjätason paikkatieto. Aikajaksolta 6006 tekstiä on ilman pitäjätietoa. Kuva: FILTER / Maciej Janicki ja Kati Kallio 2023.

puolivälissä aineistoa alkoi kertyä myös Karjalan kannakselta ja Inkeristä, ja 1860-luvulta lähtien enemmän myös läntisen Suomen alueelta. (Kuvat 8 ja 9.)

Virossa 1800-luvun alkupuolen toiminta oli baltiansaksalaisen eliitin käsissä, johon alkoi vuosisadan jälkipuoliskolla nousta myös viroa äidinkielenään puhuvia toimijoita. Vuonna 1838 perustettiin *Õpetatud Eesti Selts* (*Gelehrte Estnische Gesellschaft*) eli *Viron oppineiden seura*, joka monessa mielessä vastasi vuonna 1831 perustettua *Suomalaisen Kirjallisuuden Seuraa*. Kuten Suomessa, kansanrunoudesta keskusteltiin pääosin kirjallisuuden näkökulmasta. Keskeinen toimija oli myös kansanrunouden keräämiseen keskittynyt *Eesti Kirjameeste Selts* (1871–1893, ks. Mälk 1963). (1800-luvun kehityskuluista laajemmin ks. Kuutma 2006, 56–81; Valk 2010; 2014.) *Viron oppineiden seuran* käsikirjoituskokoelmaa ei ole vielä työstetty mukaan ERAB-korpukseen, vaikka korpuksen tekeminen muuten on aloitettu arkiston vanhimmista aineistoista. Korpus ei myöskään toistaiseksi sisällä muissa arkistoissa olevia tai ainoastaan julkaisuina säilyneitä tekstejä, kuten Alexander Heinrich Neusin kansanlaulujulkaisua vuodelta 1850.

1800-luvun loppupuolella tallentamiseen alkoivat sekä Virossa että Suomessa vaikuttaa paitsi kansakuntien rakentamisen, myös syntymässä olevan vertailevan kansanrunoudentutkimuksen tarpeet (ks. esim. Hautala 1954; Laugaste 1963; Valk 2010; 2014.) Virossa laajempi tallentaminen alkoi Jakob Hurtin sanomalehtien kautta aloittaman maanlaajuisen keruukampanjan myötä vuonna 1888. Keruuhanke herätti laajaa vastakaikua, samoin Matthias Johann Eisenin kampanja hieman myöhemmin. (Sarv ja Oras 2020, 108–109; ks. myös



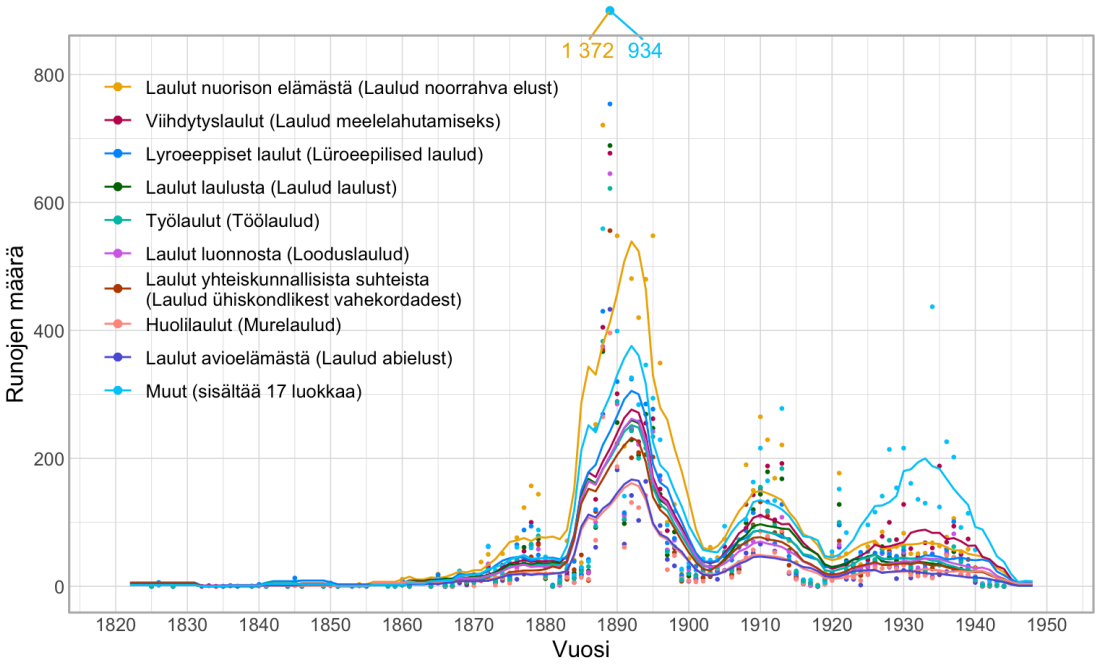
Kuva 10. Tallennetut tekstit ERAB-, JR- ja SKVR-korpuksissa vuosittain (pallot) ja kymmenen ympäröivän vuoden keskiarvona (viiva). Vasemmalla reunalla ennen vuotta 1800 kuhunkin aineistoon kertyneiden tekstien määrä. Kuvaajan yläreunasta rajautuvat pois vuosi 1889 ERAB:sta (11 785 tekstiä) sekä vuosi 1936 JR:stä (7 640 tekstiä). Kuva: Eetu Mäkelä 2023.

Hagu, Järv ja Laugaste 1989; Kuutma ja Jaago 2005.) Virolainen aineisto alkoi siis tallentua suuremmassa määrin aikana, jona tallentamisen ja vertailevan kansanrunoustieteen standardit olivat jo tarkentuneet, ja esimerkiksi paikkakunta- ja laulajatietojen merkitseminen oli säännöllisempää kuin aiemmin.

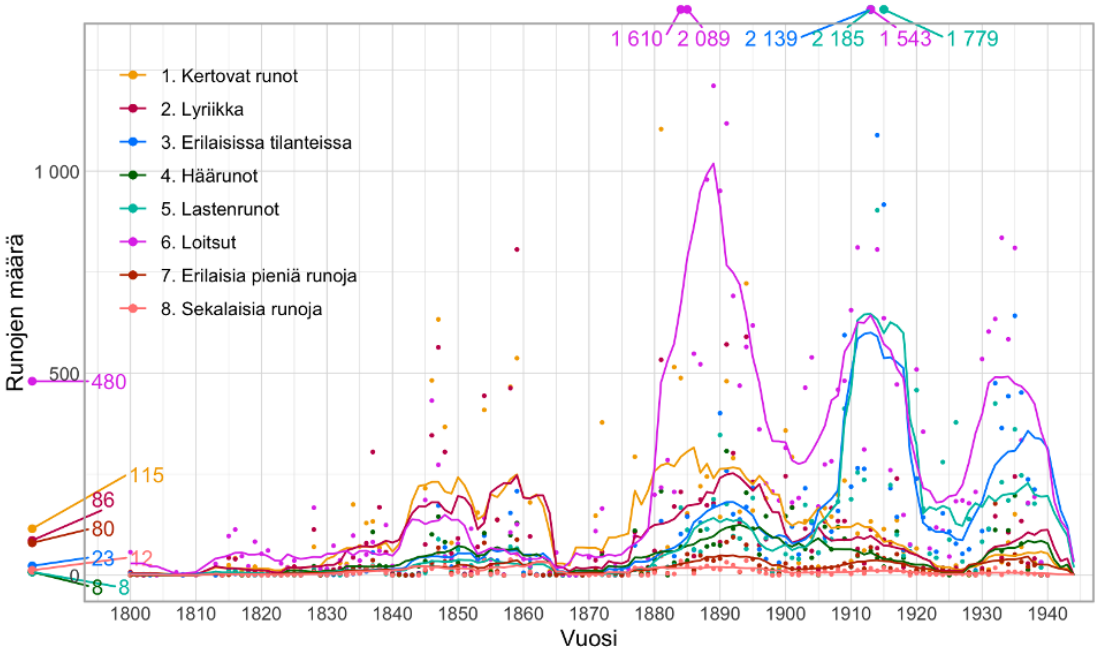
1900-luvun alusta lähtien pyrittiin huomioimaan myös aiemmin vähälle tallennukselle jääneitä alueita, ja kokoelmia karttui aiempaa tasaisemmin lähes koko itämerensuomalaiselta alueelta. Aikakauden länsisuomalaisista aineistoista melko suuri osa on lyhyehköjä loitsuja ja lastenlauluja (ks. Kuvat 12 ja 13 sivuilla 74 ja 76). Samaan aikaan kokoelmien syntyä määrittivät myös valtakuntien rajoihin ja tutkijoiden kenttätyömahdollisuuksiin vaikuttaneet muutokset – Suomen ja Viron itsenäistyminen, Inkerin läntisimpien osien kuuluminen itsenäiseen Viroon sekä Viron miehitys ja Suomen alueluovutukset (Raja-Karjala ja Karjalan kannas) toisen maailmansodan seurauksena.

Vaikka virolaisen ja suomalaisen aineiston tallennushistoriassa on erilaisia painotuksia, ajoittuvat kolme tallennusmäärien huippukohtaa samankaltaisesti 1880–1890 ja 1910–1920 lukujen taitteisiin sekä 1930-luvulle (Kuva 10).

Suurin yksittäinen vuosittainen määrä, 11 785 tekstiä, tallentui melko tasaisesti ympäri Viroa vuonna 1889 pääosin Jakob Hurtin laajan keruukampanjan seurauksena. Suomessa huippukohta on SKVR-aineiston osalta 6054 tekstiä vuonna 1913 ja JR-korpuksessa 7640 tekstiä vuonna 1936. Vuoden 1913 aineistokertymästä valtaosa on Varsinais-Suomesta ja Hämeestä (687 tekstiä Uskelasta, josta lisäksi 707 tekstiä SKVR:ssa), JR-korpuksen huippuvuodessa 1936 taas painottuvat Itä-Suomi, Raja-Karjala, Karjalan kannas sekä Viroon kuulunut Inkerin läntisin osa. Yleisesti ottaen 1900-luvun alun vuosikohtaisia tallennusprofiileja tarkasteltaessa tulee selväksi keruun painopisteen siirtyminen aiemmin niukasti tallennetun läntisen



Kuva 11. ERAB-korpuksen aineiston karttuminen runotyyppihakemiston pääluokittain. Ryhmä "Muut" sisältää tässä 17 pienempää pääluokkaa. Systematisoimattomat tyyppinimet (ks. Kuva 3) eivät kuvaajassa näy. Kuva: Eetu Mäkelä 2023.



Kuva 12. SKVR-korpuksen aineiston karttuminen runotyyppihakemiston pääluokittain. Ryhmä "Erilaisissa tilanteissa" sisältää kalendaari- ja juhlarunojen ohella myös esimerkiksi kehto- ja työlauluja. Kuva: Eetu Mäkelä 2023.

Suomen alueelle, 1930-luvulla taas tallennettiin paljon myös Viron Inkerissä, Kannaksella ja Raja-Karjalassa.

Aineiston karttuminen suhteessa runotyyppihakemistoihin näyttää paikoin liittyvän tallennuksen painottumiseen tietyille alueille sekä tiettyihin lajeihin ja aihepiireihin (kuvat 11 ja 12 edellisellä sivulla). Virolainen tallentaminen on ollut yleisprofiililtaan suomalaista tasaisempaa suhteessa sekä lajillisiin että alueellisiin jakaumiin. Paljon tallennetun, erilaisia nuorison keskenään käyttämiä lauluja sisältävän luokan "Laulut nuorison elämästä" suhteellisen osuuden laskeminen ajan myötä on muutoksista näkyvin. Lisäksi 17 pienempää pääluokkaa kuvaajassa yhdistävän Muut-kategorian osuus nousee ajan myötä. Kuvaajaa katsoessa on hyvä muistaa, että siinä näkyvät ainoastaan systematisoituun runotyyppihakemistoon liitetyt aineistot, jossa esimerkiksi häälaulujen (*pulmalaulud*, kuvaajassa osana Muut-kategoriaa) osuus on vielä melko pieni.

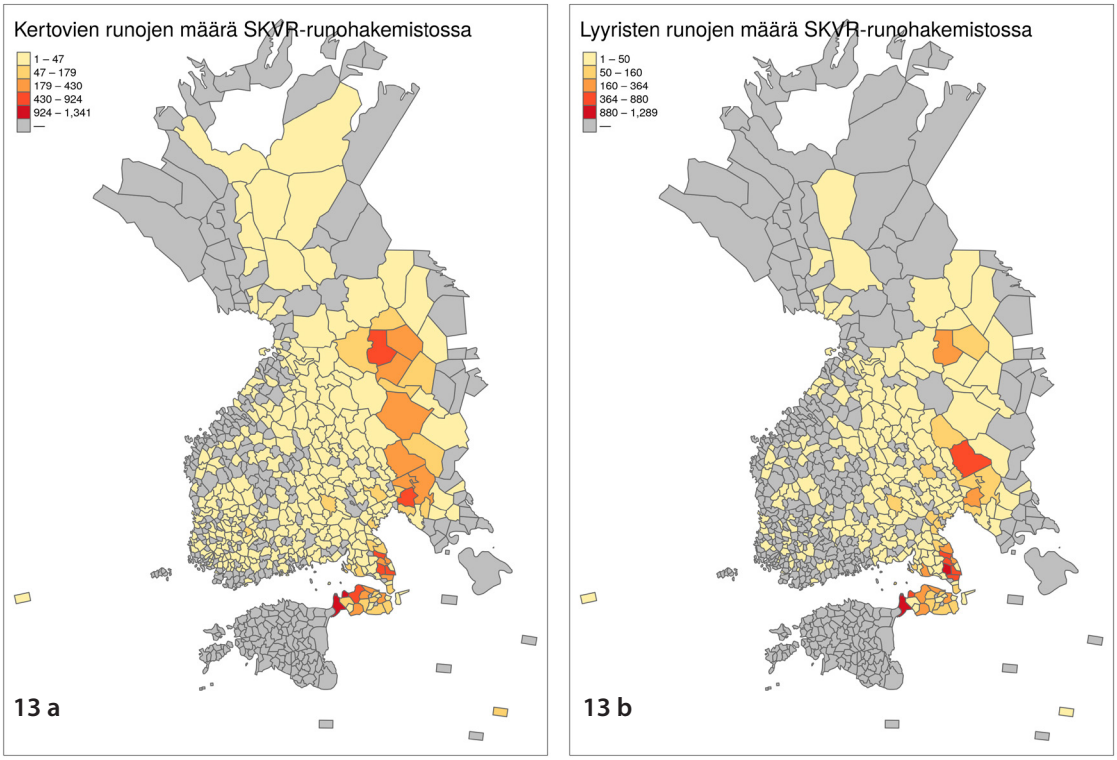
SKVR-korpuksen tyyppiluettelon ajallinen kuvaaja varioi enemmän. 1800-luvun alun tallennusta dominoi loitsuaineisto, jonka rinnalle 1830-luvulle tultaessa nousevat kertovat runot ja lyriikka. Yksittäisinä vuosina on kertynyt paljon myös yläkategoriaan "Erilaisissa tilanteissa" luokiteltuja tekstejä – tämä luokka sisältää esimerkiksi kehtolauluja, työlauluja, tanssilauluja ja kalendaarilauluja, joten sitä olisi jatkossa mielekästä tarkastella yksityiskohtaisemmin. 1800-luvun lopulla korostuvat jälleen loitsut, jotka saavat erityisesti 1910-luvulla rinnalleen lastenrunot ja luokan "Erilaisissa tilanteissa".

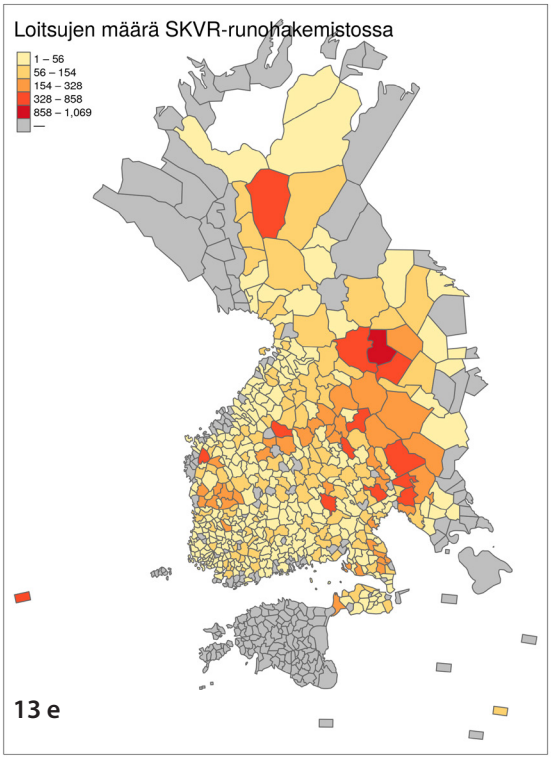
Lajien määriin kuvaajissa kuitenkin vaikuttaa myös se, miten mikäkin laji on runotyyppihakemistossa analysoitu. Esimerkiksi pienienkin motiivien tasolla SKVR-hakemistoon analysoitu lyriikka näyttäytyy kuvaajissa runsaampana kuin isoina kokonaisuuksina analysoitu epiikka.

Tarkemmat lajikohtaiset SKVR-korpuksen kartat (kuvat 13 a–e seuraavalla sivulla) näyttävät, miten epätasaisesti runotyyppijakaumat asettuvat alueellisesti. Suomalaisen aineiston epätasaisuus vaikuttaa liittyvän ennen kaikkea siihen, mitä lajeja ja runotyyppejä milläkin alueella oli tallennusaikana käytössä ja missä määrin nämä eri aikoina tallentajia kiinnostivat. Esimerkiksi kalevalamittaisia häälauluja oli luterilaisen Suomen alueella käytössä niukalti, ja tallentuneet runot ovat tyypillisesti lyhyitä eivätkä rituaalisesti keskeisiä. Kertovia ja lyyrisiä runoja tallennettiin eniten ja moninaisimpina karjalan ja inkeroisen kielen historiallisilta vaikutusalueilta, ja enemmän Itä- kuin Länsi-Suomesta. Loitsuja ja lastenlauluja taidettiin kaikkialla ja etenkin lyhyet versiot pysyivät käytössä pitkään myös Länsi-Suomessa.

Karttakuvissa 14 a–d sivulla 78 näkyvät aineiston erilaiset määrälliset ja laadulliset alueelliset painotukset. Suomeen eniten tekstejä (Kuva 14 a on tallennettu karjalankieliseltä alueelta ja Inkeristä, mutta etenkin koululaiskeruiden tuloksena myös yksittäisistä suomenkielisen alueen pitäjistä, kuten Saarijärveltä, Uskelasta ja Tyrväältä. Virossa jakauma on tasaisempi, eniten on tallennettu Pohjois- ja Etelä-Virosta ja vähiten Länsi- ja Luoteis-Virosta. Pitäjistä korostuvat määrällisesti Narvusi Länsi-Inkeristä, Kuusalu Pohjois-Virosta ja Setomaa Kaakkois-Virosta, joista kaikista on aineistossa yli 4000 tekstiä, sekä Haljala Pohjois-Virosta, Vuokkiniemi Länsi-Vienasta, Suistamo Raja-Karjalasta sekä Lempaala Karjalan kannakselta, joista on yli 3000 tekstiä kustakin.

Pitäjistä, joista on tallennettu määrällisesti eniten tekstejä, on yleisesti ottaen tallennettu myös eniten runosäkeitä (Kuva 14 b sivulla 77). Setomaalta, Narvusista ja Vuokkiniemestä

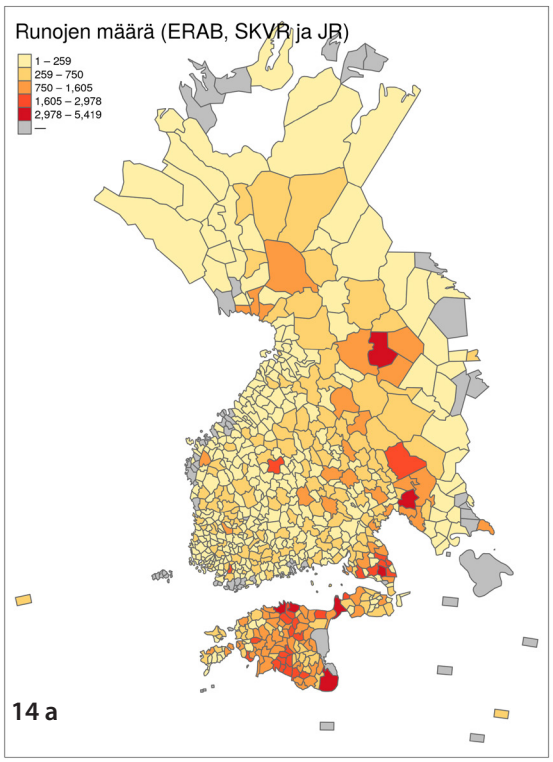

Kertovien runojen määrä SKVR-runohakemistossa
1 – 47
47 – 179
179 – 430
430 – 924
924 – 1,341
—
13 a
Lyyristen runojen määrä SKVR-runohakemistossa
1 – 50
50 – 160
160 – 364
364 – 880
880 – 1,289
—
13 b
Häärunojen määrä SKVR-runohakemistossa
1.0 – 24.0
24.0 – 67.5
67.5 – 136.5
136.5 – 394.5
394.5 – 575.0
—
13 c
Lastenrunojen määrä SKVR-runohakemistossa
1.0 – 26.5
26.5 – 79.0
79.0 – 309.0
309.0 – 646.5
646.5 – 774.0
—
13 d

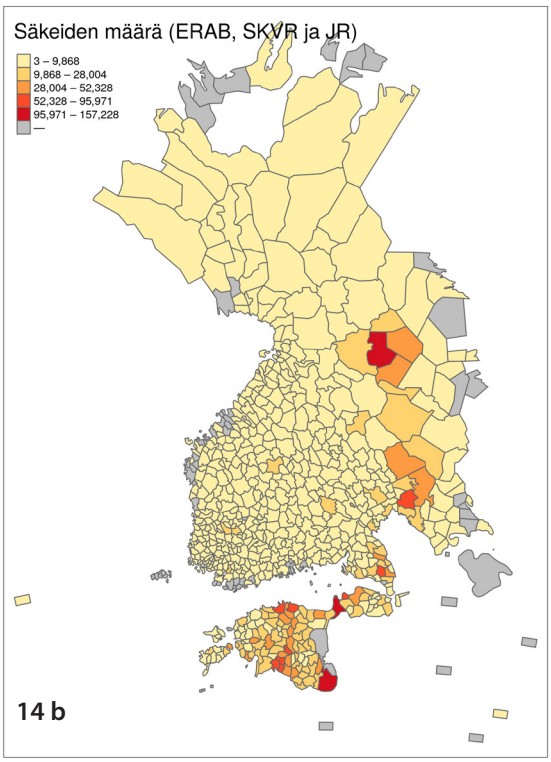


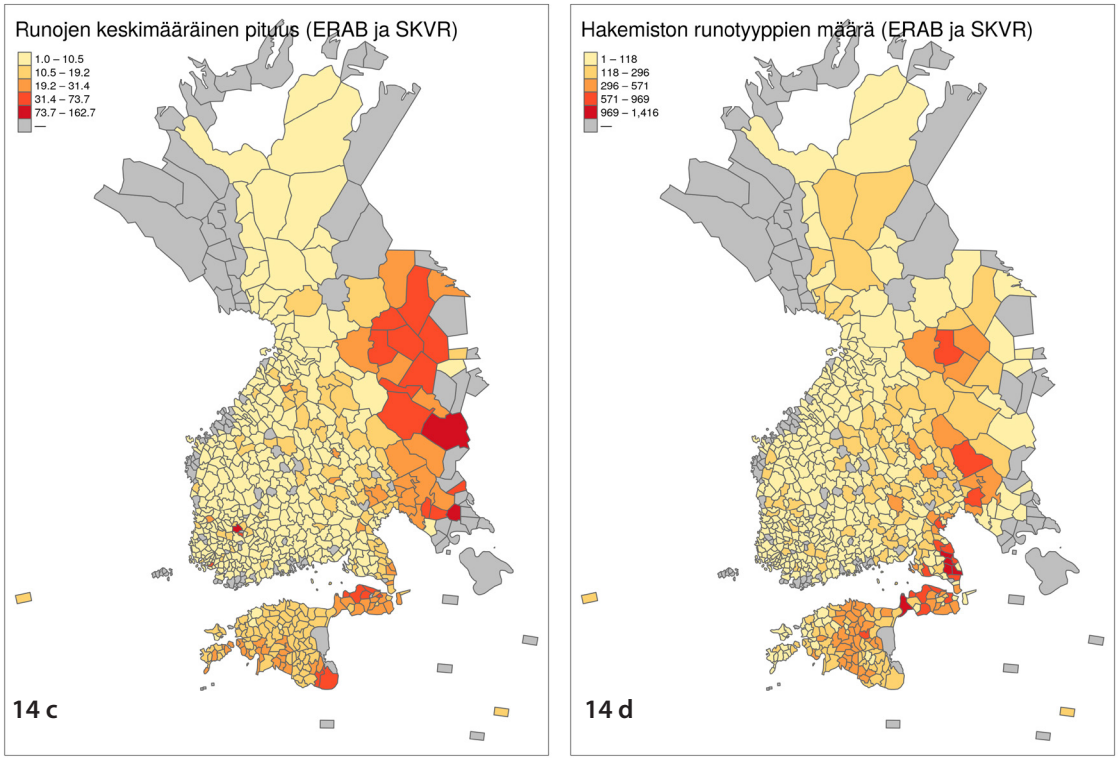
Kuva 13 a, b, c, d, e. Kertovien runojen, lyriikan, häärunojen, lastenrunojen ja loitsujen määrä pitäjittäin SKVR-runotyyppihakemistossa. Kuvat: FILTER / Maciej Janicki ja Kati Kallio 2023.

on tallennettu yli 100 000 säettä, Kuusalusta, Soikkolasta, Haljalasta, Karksista, Suistamolta, Paistusta, Lempaalasta ja Kolga-Jaanista yli 60 000 säettä kustakin.

Tallentajat näyttävät suunnanneen eniten alueille, joilla taidettiin pisimpiä runoja (Kuva 14 c). SKVR- ja ERAB-runotyyppihakemistojen pohjalta (Kuva 14 d) voi päätellä, että nämä olivat yleensä myös seutuja, joilla taidettiin eniten erilaisia lauluja ja aihelmia.







Kuva 14 a, b, c, d. Tallennettu aineisto kartalla: a) tekstien määrä, b) säkeiden määrä, c) tallennettujen runojen keskimääräinen pituus sekä d) tallentuneiden runotyyppien määrä pitäjittäin. Kaksi ensimmäistä koko aineistosta, jälkimmäiset kaksi vain ERAB- ja SKVR-korpuksista. Maakuntien tasolle tai kaupunkeihin osoitetut tai ilman tallennuspaikkatietoja jääneet runot eivät näy kartalla. Kuvat: FILTER / Maciej Janicki ja Kati Kallio 2023

Runotyyppihakemistojen pohjalta tehdyissä tarkasteluissa on kuitenkin syytä muistaa, että sekä suomalainen ja virolainen hakemisto että hakemistojen alakategoriat eroavat laatimisperiaatteiltaan ja kattavuudeltaan, mikä vaikuttaa myös määrällisiin tarkasteluihin.

ERAB	
Rosenstrauch, Karl Voldemar	3 825
Viljak, Karl	3 225
Ostrov, Mihkel	2 300
Vilberg (Vilbaste), Gustav	2 218
Viidalepp (Viidebaum), Richard	2 134
Kallas, Oskar Philipp	1 723
Seen, Gustav	1 473
Penna, Peeter	1 312
Tampere, Herbert	1 188
koguja teadmata (‘tuntematon kerääjä’)	1 175
SKVR	
Krohn, Kaarle	4 110
Paulaharju, Samuli ja Jenny	3 157
Alava, Vihtori	3 089
Europaeus, D. E. D.	2 899
Neovius, A. D.	2 535
Porkka, Volmari	2 424
Lönnrot, Elias	2 402

Perä-Pohjolan ja Lapin Kotiseutuyhdistys	2 172
Salminen, Väinö	1 761
Vihervaara, Eemeli	1 702
JR	
Hämeenlinnan alakansakouluseminaari	6 782
Perä-Pohjolan ja Lapin Kotiseutuyhdistys	3 463
Kärki, Frans	3 157
Railonsala, Artturi	2 452
Paulaharju, Samuli ja Jenny	2 168
Sääski, Sylvi	1 666
Saarijärven yhteiskoulu	1 652
Pennanen, Olavi	1 448
Paavolainen, Oma Martti	1 265
Lönnrot, Elias	1 239

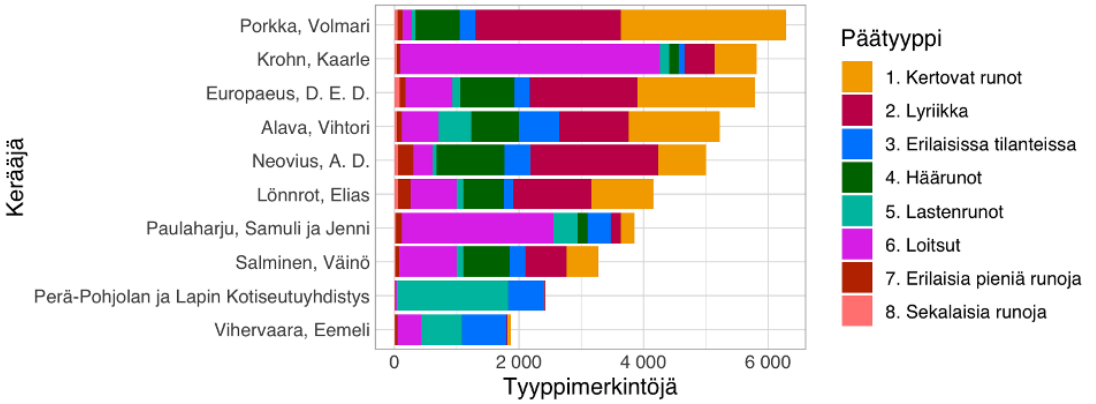
Kuva 15. Eniten tekstejä tallentaneet toimijat ERAB-, SKVR- ja JR-korpuksissa. Taulukko: Eetu Mäkelä 2023.

Yksittäisistä tallentajista (Kuva 15) tuotteliaimpia ovat ERAB-korpuksen osalta olleet Karl Voldemar Rosenstrauch (3 825 tekstiä) ja Karl Viljak (3 225), SKVR-korpuksen osalta Kaarle Krohn (4110) sekä Samuli ja Jenny Paulaharju (3157), julkaisemattomien runojen osalta taas Hämeenlinnan alakansakouluseminaari (6782) sekä Perä-Pohjolan ja Lapin kotiseutuyhdistys (3463). Koska normaalistetut tallentajasignumit ovat korpuskohtaisia, tallentajatietoja ei voi yhdistää automaattisesti. Tallentajista ainakin A. O. Väisäsellä on keräelmiä niin SKVR-, JR- kuin ERAB-korpuksessakin.

Yksittäisen tallentajan toiminta tietyllä alueella on saattanut vaikuttaa vahvasti pitäjäkohtaisten tallenteiden määrään, ajankohtaan tai painottumiseen tiettyihin lajeihin tai runotyyppeihin. Esimerkiksi Kaarle Krohnin ja Paulaharjujen tallennuksissa korostuvat yllättävänkin vahvasti loitsut, Perä-Pohjolan ja Lapin Kotiseutuyhdistyksen tallennuksissa taas lastenrunot (Kuva 16). Pohjoisessa Kittilän, Rovaniemen, Turtolan, Alatornion, Tervolan, Kemin ja Ranuan suurehkot toisintomäärät (Kuva 14 a) taas paljastuvat aineistoa yksityiskohtaisemmin tarkasteltaessa yksittäisten toimijoiden, etenkin koulukeruita organisoineen Peräpohjolan ja Lapin kotiseutuyhdistyksen tallentamiksi, ja ne koostuvat SKVR-hakemistoon indeksoidulta osaltaan ennen kaikkea lastenrunoista, kehtolauluista ja lyhyistä loitsuista.

Eri tallentajilla ja eri aikoina kiinnostus painottui siis paitsi eri seuduille, myös eri lajeihin (Kuva 16 seuraavalla sivulla). Tämä on ollut sidoksissa paitsi vaihtuviin keruu- ja tutkimusintresseihin, myös siihen, miten suosittuja ja monimuotoisia eri lajit ja runotyypit ovat eri alueilla eri aikoina olleet, mille alueille eri aikoina on ollut helppo päästä tai mitkä alueet tai lajit on eri aikoina katsottu jo riittävän hyvin tallennetuiksi (ks. myös Väisänen 1917; Haavio 1931; Hautala 1954; Kuutma ja Jaago 2005).

Vienan Karjalasta, Raja-Karjalasta, Karjalan kannaksen inkeroispitäjistä ja Inkeristä on tallennettu erityisen paljon kertovia runoja, jotka selvästi olivat näillä alueilla tallennusaikana yleisempiä, runsaampia ja pidempiä kuin muualla. Läntisen Suomen aineistoissa taas monin paikoin korostuvat lastenlaulut, kehtolaulut ja lyhyet loitsut, joita oli 1800-luvun lopun ja 1900-luvun alun paikalliskulttuureissa vielä käytöissä. Etenkin Vienasta ja Inkerin inkeroistaustaisilta paikkakunnilta on suhteessa vähemmän lastenlauluja ja kehtolauluja: tallentajat



Kuva 16. Eniten tallennuksia tehneiden toimijoiden keräelmät SKVR-aineistossa suhteessa runotyyppihakemiston pääluokkiin. Yhdellä tekstillä voi hakemistossa olla useita runotyyppinimiä: palkkien pituudet eivät tässä kuvaajassa tarkoita kunkin kerääjän tallentaman aineiston kokonaismäärää, vaan kerääjän aineistolle annettujen runotyyppiotsikoiden kokonaismäärää. Taulukko: Eetu Mäkelä 2023.

ilmeisesti keskittyivät arvokkaampina pitämiinsä lajeihin kuten kertoviin, lyyrisiin ja rituaalisiin runoihin, joita oli näillä seuduin runsaasti tallennettavissa.

1800-luvulla ja aina vuoteen 1917 asti suomalaisilla oli pääsy koko Venäjän keisarikunnan alueelle. Itsenäistymisen jälkeen huomio kiinnittyi laajemmin suomenkielisiin alueisiin, mutta myös Raja-Karjalaan, 1930-luvulla Viroon kuuluneeseen läntisimmän Inkerin alueeseen, karjalankielisiin pakolaisiin sekä toisen maailmansodan jälkeen evakoihin. Pakolaisilta ja evakoilta tallennetut aineistot on yleensä merkitty heidän alkuperäisiin kotipitäjiinsä, mutta JR-korpuksen kohdalla tässä on vaihtelua. ERAB-korpuksessa muualta kuin nykyisen Viron alueella tallennettuja aineistoja ei ole eroteltu, vaan ne on merkitty yhteen luokkaan (*välismaa*, ‘ulkomaa’).

Aineiston ja sen metatietojen ongelmia

Aineiston työstäminen FILTER hankkeessa on paljastanut yllättäviäkin tietoja itse aineistosta ja toisaalta määrällisiä varmistuksia aiemmille oletuksille. Runoregi-liittymän pohjana olevan samankaltaisuuslaskennan pohjalta esimerkiksi selvisi, että aineistossa on enemmän hyvin samanlaisia runoja kuin tiesimme ennalta. Tälle on useita syitä. Osa hyvin samankaltaisten runojen ryhmistä johtuu selvästikin suullisen perinteen laulaja-, suku- tai paikkakuntakohtaisista piirteistä: etenkin lyhyemmät runot voivat olla yllättävänkin kiinteämuotoisia. Osa samankaltaisuuksista näyttää johtuvan julkaistujen runojen, esimerkiksi D. E. D. Europaeuksen *Pieni Runon-Seppä* -kirjasen (1847), Alexander Heinrich Neusin *Ehstnische Volkslieder* -kokoelman (1850) ja Elias Lönnrotin *Kantelettaren* (1840) vaikutuksesta suulliseen kulttuuriin. Osassa tapauksia kyse on aikanaan käsikirjoitusversioina henkilöltä toiselle siirtyneestä runosta, joka on päätynyt arkistoon useana versiona osana eri keräelmiä – joskus kirjoitusasultaan tai sisällöltäänkin muokattuna. Etenkin osa kansankerääjistä on myös voinut kopioida runoja aiemmista lähteistä. Sama runo tai sen osa voi siten sisältyä erilaisina kopioina sekä SKVR- että JR-korpukseen.

Esimerkiksi Jaakko Länkelä luovutti Inkerin tallenteistaan arkistoon itse puhtaaksi kirjoittamansa version, jossa hän oli toimittanut ja yhdistellyt runoja (SKVR III1 83–452). Alkuperäiset

muistiinpanot löydettiin myöhemmin ja julkaistiin aiemmasta eroavin osin (SKVR III3 3579–3582; ks. myös SKVR III3, esipuhe). Christfrid Gananderin sanakirjan ja *Mythologica Fennica* -teoksen esimerkkisäkeitä taas julkaistiin ensin etenkin Pohjois-Pohjanmaan osissa (SKVR XII1–2), ja myöhemmin Matti Kuusi teki niistä ja muusta myöhemmin tallennetusta runoaineistosta laajempia ja ongelmallisen hypoteettisia rekonstruktioita SKVR-sarjan vuonna 1997 ilmestyneeseen täydennysosaan (SKVR XV, ks. erit. esipuhe). Tämänkaltaiset tiedot löytyvät yleensä SKVR-sarjan nidekohtaisista esipuheista, sähköisessä korpuksessa niitä ei ole. Aineistossa on myös joitain esimerkiksi sanomalehdissä julkaistuja, todennäköisesti toimitettuja runoja ja niiden mahdollisia alkuperäiskäsikirjoituksia. JR-korpuksessa on erilaisia kopioita ja kaksoiskappaleita runsaammin kuin SKVR-aineistossa.

SKVR-korpuksessa on välillä julkaistu yksittäinenkin sama runo tai runon osa tarkoituksella useaan kertaan. Joskus on haluttu julkaista runo useammassa aluekohtaisessa osassa, esimerkiksi jos runon tallennuspaikka ei ole ollut varma tai jos aluejako on hieman varioinut. Runotyyppien mukaan järjestetyissä osissa taas on paikoin haluttu julkaista useampaa runotyyppiä sisältävä runo tai sen osa useammassa kohtaa. Siten on yhdestä ja samasta runotekstistä voi yksin SKVR-aineistossa olla parhaimmillaan 11 eri pituista versiota. Joskus runoa taas ei ole julkaistu kokonaisena ollenkaan, vaan ainoastaan paloina, ja osa runosta on voinut jäädä vain käsikirjoitukseksi. Joskus SKVR-aineiston tekstiyksikkö sisältää ainoastaan viitteen toisessa kohtaa julkaistuun runoon. Lähilukuun perustuvissakin tutkimuksissa on joskus erehdytty pitämään SKVR-kirjasarjan eri kohdissa julkaistuja saman runotoisinnon versioita eri toisintoina. Vielä helpommin näin voi käydä, jos tarkastellaan esimerkiksi vain aineiston laskennallisia ominaisuuksia tai sijoittumista kartalle eikä lueta itse tekstiä.

Joskus samasta alkuperäismuistiinpanosta tai sen kopioista juontuvat tekstit on aineistossa voitu merkitä eri paikkakunnille, eri tallennusvuosille tai eri runotyyppeihin. SKVR-aineistossa osa paikkatiedoista on kirjasarjan osakohtaisten toimittajien vaihdellen päättelemiä. JR-korpuksessa paikkatiedot on yleensä pyritty merkitsemään arkistosidoksessa olevan tiedon mukaan, mutta joskus käytännöt vaihtelevat, ilmeisesti kopioijastakin riippuen. Kaikissa kolmessa osakorpuksessa paikkatiedot voivat viitata joko tekstin tallennuspaikkaan, sen esittäjän synnyin- tai asuinpaikkaan tai perinteen oppimispaikkaan. SKVR-aineistossa systematisoituja paikkakoodeja voi olla vain yksi, muita paikkatietoja voi olla vapaamuotoisessa metatietokentässä. ERAB- ja JR-korpuksissa niitä voi sitä vastoin yhdellä tekstillä olla useita. Jos paikkatietoja on useita, runot asettuvat karttaprojektioissa useampaan pitäjään ja luovat siten kuvaa todellisuutta suuremmasta aineistosta. Suurin osa aineistosta sisältää pitäjätason paikkatiedon, mutta osa vain maakuntatason, ja osa ei paikkatietoa ollenkaan (Kuva 17).

	ERAB	**SKVR**	**JR**
vain maakuntatieto	5377	3464	3130
ei paikkatietoa	3167	557	1594
ei tallennusvuotta	6696	469	0

Kuva 17. Tekstit, joille on annettu vain maakuntatason paikkatieto tai ei paikkatietoa ollenkaan tai joiden tallennusvuosi on tuntematon. ERAB-korpuksessa välismaa 'ulkomaa' (1840 tekstiä) on myös maakuntatason luokka. Luvut eivät sisällä esim. hakasulkeilla tai kysymysmerkeillä epävarmoiksi ilmaistuja tietoja. Taulukko: Maciej Janicki 2023.

Tallennusvuoden kohdalla kyse on yleensä tunnetusta tallennusvuodesta tai vuodesta, jolloin keräelmä on luovutettu arkistoon. Jos runon kopiot ovat tulleet arkistoon eri keräelmissä, myös niiden vuosimerkinnät voivat olla erilaiset. Tallennusvuosi on tuntematon ERAB-korpuksessa 6696 tekstillä ja SKVR-aineistossa 469 tekstillä – nämä on alkuperäisaineistossa merkitty eri koodeilla, vuoden 0 tai 9999 kohdalle. ERAB-korpuksessa 26 tekstillä ei ole vuosikoodia ollenkaan. Lisäksi SKVR-korpuksessa aineistoa on paikoin ajoitettu vuosisadan tarkkuudella, jolloin se on tietokantaa varten normaalistettu vuosisadan ensimmäiselle vuodelle: esimerkiksi "1600" voi tarkoittaa joko tarkkaa vuotta tai 1600-lukua. Näin ajoitetut tekstit näkyvät siis kunkin vuosisadan alkuvuoden kohdalla. Osa annetuista vuosiluvuista on aineistojen toimittajien päättelemiä tai muuten epävarmoja, ja merkitsemiskäytännöt voivat vaihdella aineistojen eri osissa.

Kerääjänimitietoja on aineistossa monenlaisia. SKVR- ja JR-korpuksissa tallentaja-kenttä viittaa kerääjänimeen SKS:n arkiston pääkortistossa. Kerääjä voi siten olla myös koulu, yhdistys tai muu vastaava toimija. Esimerkiksi Hämeen alakansakouluseminaarilta on huomattavasti aineistoa lähes koko Suomen, siihen kuuluneen Raja-Karjalan ja Karjalan kannaksen sekä Aunuksenkin alueelta vuosilta 1923–1946, Lapin ja Perä-Pohjolan kotiseutuyhdistyksellä Länsipohjan ja Peräpohjolan alueilta vuosilta 1914–1936. Tällöin yksittäiset tallentajat on usein mainittu lisätietokentässä tai joskus signumissa kerääjänimen perässä. Sanomalehdissä ilmestyneiden runojen kohdalla on käytetty niin kirjoittajan kuin lehdenkin nimeä. Joskus yhden henkilönkin keräelmässä voi olla useampien henkilöiden tallenteita. Esimerkiksi Elias Lönnrotin ja K. A. Gottlundin keräelmissä on myös heidän haltuunsa saamia vanhempia, 1700-luvulta peräisin olevia käsikirjoituksia.

Hyvin samankaltaisetkin runot – erityisesti eri alueilla esiintyvät lyhyemmät runot tai fragmentit – tai jopa saman runomuistiinpanon eri versiot on joskus voitu indeksoida eri tavoin runotyyppihakemistoon. Yksi hakemiston runotyyppi sisältää tyypillisesti sekä hyvin pitkiä että vain säkeen tai parin pituisia tekstejä. Hakemistossa asteriskilla (*) merkitty runotyyppinimi tarkoittaa, että teksti sisältää vain erityisen pienen tai tulkinnanvaraisen viitteen kyseiseen runotyyppiin. Jos hakemiston pohjalta lähtee tekemään visualisointeja, kuten karttoja tai kuvaajia, on sen luonne pidettävä mielessä. Esimerkiksi *Sammon ryöstö* ja *Sammon taonta* -runojen osumat kartalla eivät vielä kerro, että suomenkielisiltä alueilta ja ylimalkaan Vienan ja Aunuksen ulkopuolelta tallennetut tekstit ovat pääosin hyvin lyhyitä fragmentteja tai motiiveja. Irrallisena ja vaihtelevissa konteksteissa esiintyvä motiivi valtavasta lentävästä kokkolinnusta on yleensä indeksoitu *Sammon ryöstöksi*, vaikka motiivi yhdistyy yhtä lailla myös häärunoihin ja loitsuihin. (Ks. Kallio ym. 2022.)

SKVR-runohakemiston keskeinen laatija ja viimeistelijä Senni Timonen (suullinen tiedonanto 2022) onkin todennut, että hakemistoa tulee pitää ennemminkin tutkijan apuvälineenä kuin analyysin lähtökohtana, saati valmiina analyysina. Hakemisto on tehty pääosin aikana ennen sähköisiä tekstihakumahdollisuuksia ja monin paikoin aiempiin kirjasarjan osakohtaisiin hakemistoihin nojaten. Yksittäinen runotyyppiotsikko ei välttämättä kata kaikkea tutkijan tiettyyn tarkoitukseen etsimää aineistoa, ja toisaalta se saattaa sisältää myös aineistoa, jota tutkija ei itse ottaisi mukaan. Runotyyppihakemiston käyttö havainnollistaviin tai laskennallisiin tarkoituksiin edellyttää siten yleensä sen sisällön tarkistamista yksittäisten tekstien tasolla.

Systemaattisia kieli- tai etnisyystietoja korpuksiin ei sisälly, ja osa alueista on ollut monikielisiä ja -etnisiä, joten kieli- tai etnoskohtaisten tutkimusten tekeminen ei aineistosta suoraan onnistu. Karjalan kielen osalta olisi aineistoa käsiteltäessä tärkeää ottaa huomioon esimerkiksi Suomen puolen vienalaiskylät Kuivajärvi ja Hietajärvi Suomussalmella ja Rimminkylä Kuhmossa (ks. Virtaranta 1972), Raja-Karjalan monikielisyys ja kielimuotojen läheisyys (Uusitupa, Koivisto ja Palander 2017) sekä karjalankielinen substraatti Pohjois-Karjalan alueilla (esim. Hakamies 1993). Inkerin osalta vaikuttavat esimerkiksi inkeroisasutus ja -substraatti Karjalan kannaksella ja Pohjois-Inkerissä (Lauerma 2004) sekä inkeroisen, vatjan ja inkerinsuomen moninaiset suhteet, lainautumiset ja substraatit (Grünthal ja Kallio 2021).

Aineistossa on myös muita kuin kalevalamittaisia tekstejä, mutta ei kattavasti tai systemaattisesti. ERAB-korpuksesta löytyy etenkin riimillisiä ja välimuotoisia lauluja, JR-korpuksesta runsaasti riimillisiä runoja sekä joitain vepsänkielisiä lauluja, itkuvirsiä ja jopa muutamia saamenkielisiä joikuja. Toisaalta esimerkiksi sananlasku- ja arvoitusaineistot eivät ole edes kalevalamittaisen perinteen osalta kattavasti edustettuina SKVR-korpuksessa, vaikka niitä siihen satunnaisesti sisältyykin. Loitsujen osalta on SKVR:ssa ja JR:ssa – toisin kuin ERAB:ssa – mukana myös joitain kalevalamitasta poikkeavia tekstejä, mutta sanattomat taiat löytyvät ainoastaan SKS:n arkiston manuaalisesta perinnelajikortistosta. Toisaalta kaikkiin aineiston yksiköihin ei sisälly runotekstiä lainkaan: mukana on esimerkiksi proosakuvauksia ja pelkkiä viitteitä tiettyihin käsikirjoituksiin tai toisaalla julkaistuihin teksteihin. Muiden lajien tunnistaminen hakemiston kautta on mahdollista vain ERAB-korpuksessa, eikä lajihakemisto tässäkään kata koko aineistoa. Kun aineistosta tehdään määrällisiä esityksiä tai laskelmia, on siis otettava myös huomioon, että sisältö ei ole pelkkää runolaulua.

Älä luota mihinkään

Itämerensuomalaisen runolaulun sähköisessä muodossa oleva lähdeaineisto ei siis ole tasaista, kaikilta osiltaan systemaattista tai kattavaa, ja siinä on yllättäviäkin historian eri vaiheissa syntyneitä painotuksia ja ongelmia. Nämä ongelmat ovat pääosin tiedossa aiemman tutkimusten ja aineiston lähdekritiikin pohjalta, mutta määrällisiä arvioita niistä on niukalti. Aineistojen sähköinen muoto antaa niiden arviointiin ja hahmottamiseen uusia mahdollisuuksia. Samaan aikaan on selvää, että aineiston ongelmat vääristävät helposti erilaisia laskentoja, visualisointeja ja tulkintoja – ja toisaalta aineiston sähköisessä prosessoinnissakin voi mennä asioita pieleen. Onkin tärkeää, että tutkijat myös sähköisiä välineitä käyttäessään yhä tuntevat aineiston taustat, lukevat yksittäisiä tekstejä, suhtautuvat aineiston metatietoihin varauksella ja osaavat tarvittaessa hakeutua alkuperäisten muistiinpanojen tai laajemman arkistoaineiston pariin. Välineiden kehittäjän näkökulmasta tekisi mieli liittää jokaiseen näkymään punaisena vilkkuvia varoitustarroja erilaisista tavoista, joilla niin itse aineisto kuin sen laskennalliset käsittelytavatkin saattavat johtaa harhaan ja virhepäätelmiin.

Sähköinen olomuoto ja mahdollisuus sanahakuihin voi saada aineiston tuntumaan aiempaa helpommalta lähestyä. Digitaalisilla välineillä haettaessa osa aineistosta jää kuitenkin helposti katveeseen tai keskeisiä aineiston tekstien välisiä yhteyksiä hahmottumatta. Osa näistä yhteyksistä – laajempi tilannekonteksti tai vaikkapa yhden tallennustilanteen aikana muodostunut temaattinen tai toisiaan kommentoivien runojen ja muiden lajien ketju – hahmottuu ainoastaan alkuperäisten käsikirjoitusten tasolla. Lisäksi runoaineistot kytkeytyvät moniin muihinkin aineellisen ja aineettoman perinteen alueisiin, joista mahdollisesti tallennettuja tietoja ei usein ole runoaineistojen itsensä yhteydessä. Käsikirjoitusten tulkintaan

tarvitaankin usein laajemman mikrohistoriallisen aineiston, muiden arkistokokoelmien, tallennushistorian tai kulttuuristen ja historiallisten kontekstien tuntemusta (ks. esim. Timonen 2004; Tarkka 2005; Oras 2008.) Onkin toivottava, että sähköisten aineistojen helppo saatavuus ei vähennä paneutuvan arkistotutkimuksen määrää.

Nämä varaukset huomioiden laskennalliset välineet kuitenkin antavat mahdollisuuksia niin aineiston kokonaisuuden, painotusten ja hiljaisuuksien arviointiin, sen erilaisten osien toisiinsa suhteuttamiseen kuin sen läpi leikkaavien monitasoisten samankaltaisuuksienkin tunnistamiseen ja analyysiin. Aineistoja tässä esitettyä yksityiskohtaisemmin tarkastelemalla olisi esimerkiksi mahdollista luoda tallennushistorian kulusta tai sen yksityiskohdista aiempaa tarkempaa ja laajempaa kokonaiskuvaa.

Tutkimusaineistot ja käyttöliittymät

ERAB-korpus: *Eesti Regilaulude Andmebaas*. 2003–. Toimittaneeet Janika Oras, Liina Saarlo, Mari Sarv, Kanni Labi, Merli Uus ja Reda Šmitaite. Tartto: Eesti Kirjandusmuuseumi Eesti Rahvaluule Arhiiv. https://www.folklore.ee/regilaul/andmebaas. (Huhtikuun 2023 versio.)

FILTER-tietokanta. 2020–. Maciej Janicki ja Eetu Mäkelä. Helsingin yliopisto, Suomalaisen Kirjallisuuden Seura ja Eesti Rahvaluule Arhiiv.

FILTER Visualizations. 2022–. Maciej Janicki sekä Kati Kallio, Eetu Mäkelä, Jukka Saarinen ja Mari Sarv. Hankkeen sisäisessä käytössä oleva demoversio. Helsingin yliopisto, Suomalaisen Kirjallisuuden Seura ja Eesti Rahvaluule Arhiiv.

JR-korpus. Julkaisemattomat runot, sähköinen versio (huhtikuu 2023). Helsinki: Suomalaisen Kirjallisuuden Seuran arkisto. Käytössä FILTER-hankkeessa tuotettu versio, josta mahdollisesti elossa olevien informanttien henkilötiedot on poistettu.

Octavo UI: Finnic Oral Poetry. 2017–. Eetu Mäkelä. https://jiemakel.github.io/octavo-nui/#/search?endpoint=https%3A%2F%2Ffilter-octavo.rahtiapp.fi%2Ffilter%2F&level=POEM. Helsingin yliopisto.

Runoregi, versio 2.0. 2022–. Maciej Janicki, Kati Kallio, Mari Sarv ja Eetu Mäkelä. Helsinki ja Tartto: Helsingin yliopisto (HELDIG), Suomalaisen Kirjallisuuden Seura ja Eesti Kirjandusmuuseum. https://runoregi.rahtiapp.fi.

SKSÄ 2023:3: Senni Timosen haastattelu 13.2.2023 Helsingissä. Haastattelijat Jukka Saarinen ja Kati Kallio. Suomalaisen Kirjallisuuden Seuran arkisto, Perinteen ja nykykulttuurin kokoelma.

SKSÄ 2023:30 Senni Timosen haastattelu 23.2.2023 Helsingissä. Haastattelijat Jukka Saarinen ja Kati Kallio. Suomalaisen Kirjallisuuden Seuran arkisto, Perinteen ja nykykulttuurin kokoelma.

SKVR-korpus, versio 2.0. (13.9.2022.). Suomalaisen Kirjallisuuden Seura. https://github.com/sks190/SKVR.

SKVR-tietokanta. 2004–. Toimittaneet Jukka Saarinen ja Arvo Krikmann. Helsinki: Suomalaisen Kirjallisuuden Seura. URN:[NBN:fi-fe20051411.], https://skvr.fi.

Timonen, Senni. 2022: Suullinen tiedonanto Kati Kalliolle (artikkelissa esitetty tulkinta tarkistettu Timoselta).

Kirjallisuus

Abondolo, Daniel ja Riitta-Liisa Valijärvi. 2023. *The Uralic Languages*. Second edition. Abingdon Oxon: Routledge. https://doi.org/10.4324/9781315625096.

Anttonen, Pertti. 2005. *Tradition through Modernity: Postmodernism and the Nation-State in Folklore Scholarship.* Helsinki: Finnish Literature Society. https://doi.org/10.21435/sff.15.

Ariste, Paul. 1960. *Vadjalaste laule.* Tallinna: Emakeele Selts.

Arkiston avain: Kansanrunousarkiston kortistot, hakemistot, luettelot ja lyhenteet. 1984. Helsinki: Suomalaisen Kirjallisuuden Seura.

Bakró-Nagy, Marianne, Johanna Laakso ja Elena Skribnik, toim. 2022. *The Oxford Guide to the Uralic Languages*. First edition. Oxford Guides to the World's Languages. New York: Oxford University Press. https://doi.org/10.1093/oso/9780198767664.001.0001.

Bauman, Richard ja Charles L. Briggs. 2003. *Voices of Modernity: Language Ideologies and the Politics of Inequality.* Cambridge: Cambridge University Press.

Bendix, Regina. 1997. *In Search of Authenticity: The Formation of Folklore Studies.* Madison (WI): University of Wisconsin Press.

Bendix, Regina ja Galit Hasan-Rokem, toim. 2012. *A Companion to Folklore*. Malden (MA): Wiley-Blackwell.

Europaeus, David Emanuel Daniel. 1847. *Pieni Runon-Seppä eli Kokous paraimmista Inkerinmaan puolelta kerätyistä runo-lauluista, ynnä Johdatuksia runon tekoon D. E. D. Europaeukselta.* Helsinki: Simeliuksen perilliset. Digitoituna: www.urn.fi/urn:nbn:fi:sks-dor-002039.

Eesti Kirjandusmuuseum. N.d. *Kirjandusmuuseum.* Luettu 25.4.2023. https://www.kirmus.ee/et/kirjandusmuuseum.

Jevsejev, V. Ja., toim. 1950. *Karelskije epitšeskije pesni.* Leningrad: Akademija nauk SSSR, Karelo-finski filial AN SSSR, Institut istorii, jazyka i literatury.

Jevsejev, V. Ja., toim. 1994. *Karelo-finskij narodnyj epos. Karjalais-suomalainen kansaneepos.* Moskva: Vostočnaâ literatura.

Grünthal, Riho ja Kati Kallio. 2021. "Inkerinmaan ja Setomaan historialliset kulttuurit." Teoksessa *Inkerikot, setot ja vatjalaiset: Kansankulttuuri, kieli ja uskomusperinteet*, toimittaneet Kati Kallio, Riho Grünthal ja Lassi Saressalo, 9–32. Helsinki: Suomalaisen Kirjallisuuden Seura. https://doi.org/10.21435/skst.1467.

Haavio, Martti. 1931. *Kansanrunouden keruu ja tutkimus.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Hagu, Paul, A. Järv ja E. Laugaste, toim. 1989. *Jakob Hurda teened rahvaluuleteaduse arendamisel.* Tartto: Tartu Riiklik Ülikool.

Hakamies, Pekka, toim. 1990. *Runo, alue, merkitys: Kirjoituksia vanhan kansanrunon alueellisesta muotoutumisesta.* Joensuu: Joensuun yliopisto.

Hakamies, Pekka. 1993. *Venäjän-Taipaleelta Viinijärvelle: Erään karjalaisryhmän identiteetistä ja assimilaatiosta.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Harvilahti, Lauri. 1992. *Kertovan runon keinot: Inkeriläisen runoepiikan tuottamisesta.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Harvilahti, Lauri. 2013. "The SKVR Database of Ancient Poems of the Finnish People in Kalevala Meter and the Semantic Kalevala." *Oral Tradition* 28(2): 223–232. Luettu 22.4.2023. http://journal.oraltradition.org/issues/28ii/harvilahti.

Harvilahti, Lauri. 2019. History of Computational Folkloristics in Finland and Some Current Perspectives. Teoksessa *Folkloristics in the Digital Age*, toimittaneet Pekka Hakamies ja Anne Heimo, 158–175. Helsinki: Suomalainen Tiedeakatemia.

Hautala, Jouko. 1950. *Suomen kansan vanhojen runojen julkaisutyön vaiheita*. Eripainos SKVR-sarjan osasta XIV. Helsinki: Suomalaisen Kirjallisuuden Seura.

Hautala, Jouko. 1954. *Suomalainen kansanrunoudentutkimus.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Hämäläinen, Niina. 2012. *Yhteinen perhe, jaetut tunteet: Lyyrisen kansanrunon tekstualisoinnin ja artikuloinnin tapoja Kalevalassa*. Turku: Turun yliopisto. https://urn.fi/URN:ISBN:978-951-29-5170-3.

Ilyefalvi, Emese. 2018. "The Theoretical, Methodological and Technical Issues of Digital Folklore Databases and Computational Folkloristics." *Acta Ethnographica Hungarica* 63(1): 209–258. https://doi.org/10.1556/022.2018.63.1.11.

Jaago, Tiiu. 1999. "Rahvaluule mõiste kujunemine Eestis." *Mäetagused* 9: 70–91. https://doi.org/doi:10.7592/MT1999.09.rhl.

Jaago, Tiiu. 2010. "Traditsioon ja regilaul folkloristika muutuvas kontekstis." *Keel ja kirjandus* 53(8): 592–610.

Janicki, Maciej. 2022. "Optimizing the Weighted Sequence Alignment Algorithm for Large-Scale Text Similarity Computation." Teoksessa *Proceedings of the 2nd International Workshop on Natural Language Processing for Digital Humanities*, toimittaneet Mika Hämäläinen, Khalid Alnajjar, Niko Partanen ja Jack Rueter, 96–100. Taipei, Taiwan: Association for Computational Linguistics. https://aclanthology.org/2022.nlp4dh-1.13. Luettu 22.4.2023.

Janicki, Maciej (tekeillä). "FILTER Database User Guide." https://a3s.fi/filter-db-doc/intro.html. Luettu 23.4.2023.

Janicki, Maciej, Kati Kallio ja Mari Sarv. 2023. Exploring Finnic Written Oral Folk Poetry through String Similarity. *Digital Scholarship in the Humanities* 38(1): 180–194. https://doi.org/10.1093/llc/fqac034.

Järv, Risto. 2013. "Estonian Folklore Archives." *Oral Tradition* 28 (2): 291–298. https://doi.org/10.1353/ort.2013.0022.

Järv, Risto. 2016. "The Singing Wolf Meets His Kin Abroad: Web-Based Databases of the Estonian Folklore Archives". *Estudis de Literatura Oral Popular. Studies in Oral Folk Literature* 5: 29–44. Luettu 22.4.2023. https://raco.cat/index.php/ELOP/article/view/327998.

Järv, Risto ja Mari Sarv. 2014. "From Regular Archives to Digital Archives." Teoksessa *Corpora ethnographica online. Strategies to digitize ethnographical collections and their presentation on the Internet. Rostocker Studien zur Volkskunde und Kulturgeschichte* 5, toimittanut Christoph Schmitt, 49–60. Münster: Waxmann Verlag GmbH.

Järvinen, Irma-Riitta. 2008. "Perspectives to the Relations between the Estonian Folklore Archives and the Folklore Archives of the Finnish Literature Society." *Journal of Ethnology and Folkloristics* 2(2): 57–67. Luettu 22.4.2023. https://www.jef.ee/index.php/journal/article/view/30.

Kalkun, Andreas. 2015. *Seto laul eesti folkloristika ajaloos: Lisandusi representatsiooniloole.* Tartto: Eesti Kirjandusmuuseumi Teaduskirjastus.

Kallio, Kati ja Eetu Mäkelä. 2019. "Suullisen runon sähköisestä lukemisesta." *Elore* 26(2): 26–41. https://doi.org/10.30666/elore.84570.

Kallio, Kati, Frog ja Mari Sarv. 2017. "What to Call the Poetic Form: Kalevala-Meter or Kalevalaic Verse, regivärss, Runosong, the Finnic Tetrameter, Finnic Alliterative Verse, or Something Else?" *RMN Newsletter* 12–13: 94–117. Luettu 22.4.2023. http://hdl.handle.net/10138/305420.

Kallio, Kati, Eetu Mäkelä ja Maciej Janicki. 2020. "Historical Oral Poems and Digital Humanities: Starting with a Finnish Corpus." *Folklore Fellows Network* 54 (1):12–18. Luettu 22.4.2023. https://www.folklorefellows.fi/historical-oral-poems-and-digital-humanities/.

Kallio, Kati, Maciej Janicki, Eetu Mäkelä ja Mari Sarv. 2022. "Recognising Intertextuality in the Digital Corpus of Finnic Oral Poetry. Experiment with the Sampo Cycle." *CEUR Workshop Proceedings* 3232: 279–87. Luettu 22.4.2023. http://ceur-ws.org/Vol-3232/paper26.pdf.

Klemettinen, Pasi, toim. 2006. *"Ei se synny synnyttämättä". Selvitys digitointiprojektin vaiheista ja työprosesseista.* Helsinki: Suomalaisen Kirjallisuuden Seura. Luettu 22.4.2023. https://www.finlit.fi/sites/default/files/mediafiles/tutkimus/elias_loppuraportti.pdf.

Krohn, Kaarle. 1916. *Suomen kansan vanhojen runojen julkaisu*. Porvoo: WSOY.

Kundozerova, Maria. 2022: "Baza dannyh "Karelskije runy": ideja sozdanija, kontseptsija, perspektivy." *Almanah severojevropejskih i baltijskih issledovanij* 7: 233–240. http://dx.doi.org/10.15393/j103.art.2022.2386.

Kurki, Tuulikki, toim. 2004. *Kansanrunousarkisto, lukijat ja tulkinnat.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Kuusi, Matti. 1977. "Viron, Inkerin ja Karjalan kansanrunovalikoimia." *Virittäjä* 81(2): 214–218. https://journal.fi/virittaja/article/view/36855.

Kuusi, Matti ja Ülo Tedre. 1979. "Regivärsilise ja kalevalamõõdulise laulutraditsiooni vahekorrast: Dialog üli lahe." *Keel ja kirjandus* 22: 70–78.

Kuutma, Kristin 2006. *Collaborative Representations: Interpreting the Creation of a Sámi Ethnography and a Seto Epic.* Helsinki: Suomalainen Tiedeakatemia.

Kuutma, Kristin. 2015. "From Folklore to Intangible Heritage." Teoksessa *A Companion to Heritage Studies*, toimittaneet William Logan, Máiréad Nic Craith ja Ullrich Kockel, 41–54. Chichester, West Sussex: Wiley-Blackwell. https://doi.org/10.1002/9781118486634.ch3.

Kuutma, Kristin ja Tiiu Jaago, toim. 2005. *Studies in Estonian Folkloristics and Ethnology. A Reader and Reflexive History.* Tartto: Tartu University Press.

Kõiva, Mare. 2019. *Eesti loitsud I. Arstimissõnad I*. Tartto: Eesti Kirjandusmuuseumi Teaduskirjastus.

Laaksonen, Pekka ja Jukka Saarinen, toim. 2004. *Arkiston avain: Kansanrunousarkiston kortistot, hakemistot, luettelot, lyhenteet.* 2. uud. laitos. Helsinki: Suomalaisen Kirjallisuuden Seura.

Laugaste, Eduard. 1963. *Eesti rahvaluuleteaduse ajalugu: Valitud tekste ja pilte.* Tallinna: Eesti riiklik kirjastus.

Laugaste, Eduard. 1975. *Eesti rahvaluule.* Tallinna: Valgus.

Lauerma, Petri. 2004. *Larin Parasken epiikan kielellisestä variaatiosta*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Lavonen, N. A., toim. 1989. *Kiestingin kansanlauluja. Pesennyj folklor kestengskih karel.* Petrozavodsk: Karelija.

Loorits, Oskar. 1932. "Eesti rahvaluuleteaduse tänapäev. Olevase ja tulevase töökava." Teoksessa *Vanavara vallast. Õpetatud Eesti Seltsi Kirjad* I, 7–34. Tartto: Õpetatud Eesti Selts.

Lönnrot, Elias. 1840. *Kanteletar taikka Suomen kansan wanhoja lauluja ja wirsiä.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Mironova, V. P., toim. 2006. *Epitšeskije pesni Južnoj Karelii. Anuksen karjalazien eeppizet pajot.* Petrozavodsk: Periodika.

Mikkola, Kati. 2021. "Vähemmistöjen roolit muuttuvassa arkistopolitiikassa. Perinnekokoelmien etnisiä ja kielellisiä rajanvetoja Suomessa ja Virossa." Teoksessa *Arkistot ja kulttuuriperintö*, toimittaneet Outi Hupaniittu ja Ulla-Maija Peltonen, 166–210. Helsinki: Suomalaisen Kirjallisuuden Seura. https://doi.org/10.21435/tl.268.

Mäkelä, Heidi Henriikka ja Lotte Tarkka. 2022. "Sopimatonta: Seksuaalisuuteen liittyvien kalevalamittaisten runojen perinnöllistäminen Suomessa 1818–1997." *Elore* 29(2): 34–58. https://doi.org/10.30666/elore.121473.

Mälk, Vaina. 1963. *Eesti Kirjameeste Seltsi osa eesti folkloristika arengus*. Tallinna: Eesti Riiklik Kirjastus.

Neus, Alexander Heinrich. 1850. *Estnische Volkslieder. Erste Abtheilung.* Reval: Kluge und Ström. Digitoitu. Luettu 23.4.2023. https://utlib.ut.ee/eeva/index.php?lang=et&do=tekst&tid=176.

Nissilä, Viljo, Matti Sarmela ja Aino Sinisalo. 1970. *Ethnologisches Ortschafts- und Dorfregister des finnischen Sprachgebiets: In Finnland, Karelien, Ingermanland, Norrbotten, Finnmark und auf der Halbinsel Kola*. Studia Fennica 15. Helsinki: Suomalaisen Kirjallisuuden Seura.

Oras, Janika. 2008. *Viie 20. sajandi naise regilaulumaailm. Arhiivitekstid, kogemused ja mälestused.* Tartto: Eesti Kirjandusmuuseumi Teaduskirjastus.

Rüütel, Ingrid ja Koit Haugas. 1990. "A Method for Distinguishing Melody Types and Establishing Typological Groups (on the Material of Estonian Runo Songs)." *Musikometrika* 2. *Quantitative Linguistics*. Vol. 43, 169–186.

Rüütel, Ingrid 1999. "Varafolkloorsetelt vokaalzhanridelt lauluni III." *Mäetagused. Hüperajakiri* 10: 90–105. https://doi.org/10.7592/MT1999.10.rtl.

Saarinen, Jukka. 2001. "Kalevalaic Poetry as a Digital Corpus." *FF Network* 22: 4–9. Luettu 23.4.2023. https://www.folklorefellows.fi/kalevalaic-poetry-as-a-digital-corpus/.

Saarinen, Jukka. 2006. "SKVR-tietokanta." Teoksessa *"Ei se synny synnyttämättä". Selvitys digitointiprojektin vaiheista ja työprosesseista*, toimittanut Pasi Klemettinen, 36–43. Helsinki: Suomalaisen Kirjallisuuden Seura.

Saarlo, Liina. 2012. "Kaks hõbekeelset kannelt helisemas taas..." *Keel ja Kirjandus* 11: 836–846. https://doi.org/10.54013/kk660a4.

Sarajas, Annamari. 1956. *Suomen kansanrunouden tuntemus 1500–1700-lukujen kirjallisuudessa.* Porvoo: WSOY.

Sarv, Mari 2008. *Loomiseks loodud: regivärsimõõt traditsiooniprotsessis.* Eesti Rahvaluule Arhiivi toimetused. Commentationes Archivi Traditionum Popularium Estoniae 26. Tartto: Eesti Kirjandusmuuseumi Teaduskirjastus.

Sarv, Mari. 2019. "Poetic Metre as a Function of Language. Linguistic Grounds for Metrical Variation in Estonian Runosongs." *Studia Metrica et Poetica* 6(2): 102–48. https://doi.org/10.12697/smp.2019.6.2.04.

Sarv, Mari. 2020. "Regilaulude teema-analüüs: võimalusi ja väljakutseid / Topic Analysis of Estonian Runosongs: Prospects and Challenges." *Methis. Studia Humaniora Estonica* 21(26): 137–60. https://doi.org/10.7592/methis.v21i26.16914.

Sarv, Mari ja Janika Oras. 2020. "From Tradition to Data: The Case of Estonian Runosong." *ARV. Nordic Yearbook of Folklore* 76: 105–117. Luettu 23.4.2023. https://gustavadolfsakademien.bokorder.se/sv-SE/serie/140/arv.

Sarv, Mari, Kati Kallio, Maciej Janicki ja Eetu Mäkelä. 2021. "Metric Variation in the Finnic Runosong Tradition: A Rough Computational Analysis of the Multilingual Corpus." Teoksessa *Tackling the Toolkit. Plotting Poetry through Computational Literary Studies*, toimittaneet Petr Plecháč, Robert Kolár, Anne-Sophie Bories ja Jakub Říha, 131–150. Prague: Institute of Czech Literature CAS. https://doi.org/10.51305/ICL.CZ.9788076580336.

Sihvo, Hannes. (1973) 2003. *Karjalan kuva. Karelianismin taustaa ja vaiheita autonomian aikana.* 2. tark. ja täyd. painos. Helsinki: Suomalaisen Kirjallisuuden Seura.

Siikala, Anna-Leena. 2000. "Body, Performance, and Agency in Kalevala Rune-Singing." *Oral Tradition* 15(2): 255–278.

SKVR: *Suomen Kansan Vanhat Runot* I–XV. 1908–1948, 1997. Useita toimittajia. Helsinki: Suomalaisen Kirjallisuuden Seura.

Stepanova, A. S. 2000. *Ustnaja poezija tungudskih karel. Tunguon rahvahan suusanallista perinnehtä.* Petrozavodsk: Periodika.

Tangherlini, Timothy R. 2016. "Big Folklore: A Special Issue on Computational Folkloristics." *The Journal of American Folklore* 129(511): 5–13. https://doi.org/10.5406/jamerfolk.129.511.0005.

Tarkka, Lotte. 1989. "Karjalan kuvaus kansallisena retoriikkana. Ajatuksia karelianismin etnografisesta asetelmasta." Teoksessa *Runon ja rajan tiellä.* Kalevalaseuran vuosikirja 68, toimittaneet Seppo Knuuttila ja Pekka Laaksonen, 243–57. Helsinki: Suomalaisen Kirjallisuuden Seura.

Tarkka, Lotte. 2005. *Rajarahvaan laulu. Tutkimus Vuokkiniemen kalevalamittaisesta runokulttuurista 1821–1921*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Tarkka, Lotte, Heidi Haapoja-Mäkelä ja Eila Stepanova. 2019. "Kalevalaisuus, kieli-ideologiat ja suomalaisuuden myytit". Teoksessa *Eurooppa, Suomi, Kalevala. Mikä mahdollisti Kalevalan?* Kalevalaseuran vuosikirja 98, toimittaneet Ulla Piela, Pekka Hakamies ja Pekka Hako, 79–106. Helsinki: Suomalaisen Kirjallisuuden Seura.

Tedre, Ülo. 1969–1974. *Eesti rahvalaulud. Antoloogia.* Tallinna: Eesti Raamat. Verkkoversio: https://www.folklore.ee/laulud/erla/indeks1.html.

Timonen, Senni. 2004. *Minä, tila, tunne. Näkökulmia kalevalamittaiseen kansanlyriikkaan*. Helsinki: Suomalaisen Kirjallisuuden Seura.

Uusitupa, Milla, Vesa Koivisto ja Marjatta Palander. 2017. "Raja-Karjalan murteet ja raja-alueiden kielimuotojen nimitykset." *Virittäjä* 121(1): 67–106. Luettu 23.4.2023. https://journal.fi/virittaja/article/view/53121.

Valk, Ülo. 2010. "Eesti folkloristika kulg distsipliinist diskursiivseks formatsiooniks." *Keel ja Kirjandus* 8–9: 561–74. Luettu 23.4.2023. https://keeljakirjandus.ee/archive/Valk%20561-574.pdf.

Valk, Ülo. 2014. "Folkloristic Contributions towards Religious Studies in Estonia: A Historical Outline." *Temenos. Nordic Journal of Comparative Religion* 50(1): 137–164. https://doi.org/10.33356/temenos.46254.

Vikis-Freibergs, Vaira ja Imants Freibergs. 1978. "Formulaic Analysis of the Computer-Accessible Corpus of Latvian Sun-Songs." *Computers and the Humanities* 12(4): 329–339.

Virtaranta, Pertti. 1972. *Polku sammui: Vienalaiskylien vaiheita rajan molemmin puolin*. Helsinki: Kirjayhtymä.

Väisänen, A. O. 1917. *Suomen kansan sävelmäin keräys: Vaiheet ja tulokset.* Helsinki: Suomalaisen Kirjallisuuden Seura.

Wilson, William A. 1976. *Folklore and Nationalism in Modern Finland*. Bloomington, IN: Indiana University Press.

FT Kati Kallio on akatemiatutkija Suomalaisen Kirjallisuuden Seurassa ja folkloristiikan dosentti Helsingin yliopistossa. Hän on perehtynyt suomalaisten arkistojen kalevalamittaisen runouden kokoelmiin ja kiinnostunut erityisesti paikallisista laulukulttuureista.

TT Maciej Janicki on post doc -tutkija ihmis- ja tietojenkäsittelytieteiden vuorovaikutuksen tutkimusryhmässä Helsingin yliopistossa. Hän on valmistunut tohtoriksi Leipzigin yliopiston tietojenkäsittelytieteestä ja on kiinnostunut erityisesti strukturoimattoman kielellisen aineiston käsittelemisestä ohjaamattomien menetelmien avulla.

TT Eetu Mäkelä on apulaisprofessori Helsingin yliopistossa ja tietojenkäsittelytieteen dosentti Aalto-yliopistossa. Hän johtaa Helsingin digitaalisten ihmistieteiden keskuksessa monitieteistä tutkimusryhmää, joka selvittää teknisiä, prosessuaalisia ja teoreettisia edellytyksiä menestyksekkäälle laskennalliselle tutkimukselle humanistisilla ja yhteiskuntatieteellisillä aloilla.

FT Jukka Saarinen on vanhempi tutkija FILTER hankkeessa ja kalevalamittaisen runouden aineistojen asiantuntija. Hän on työskennellyt Suomalaisen Kirjallisuuden Seuran arkiston kehittämispäällikkönä ja vastannut arkiston sähköisistä tietojärjestelmistä ja aineistoista.

FT Liina Saarlo on tutkija Viron Kansanrunousarkistossa Viron Kirjallisuusmuseossa. Hän on perehtynyt erityisesti virolaisen kansanrunouden aineistoihin sekä kansanrunouden formulaisuuteen ja typologioihin.

FT Mari Sarv on vanhempi tutkija Viron kansanrunousarkistossa Viron Kirjallisuusmuseossa. Hän on erikoistunut itämerensuomalaisen suullisen runouden kieleen ja poetiikkaan ja käyttänyt tutkimuksessaan myös laskennallisia menetelmiä.