MARJATTA PALANDER, HELKA RIIONHEIMO,
HANNU KEMPPANEN & JUKKA MÄKISALO
Itä-Suomen yliopisto

Kielikorpuksia Suomen itärajalta

1. Johdanto

Itä-Suomen yliopistossa on pitkät perinteet raja-alueiden ja erityisesti Suomen itärajan monitieteisessä tutkimuksessa. Yliopiston nykyisessä tutkimusstrategiassa yksi kansainvälisistä huipputason tutkimusalueista on *Rajat, liikkuvuus ja kulttuurien kohtaaminen*, joka on edustettuna paitsi yhteiskunta- ja historiatieteissä myös kieli- ja käännöstieteissä, mm. suomen ja venäjän kielessä sekä karjalan kielen ja kulttuurin oppiaineessa. Näissä kieliaineissa on viime vuosina työstetty kolmea digitaalista korpusta (*Raja-Karjalan korpus*, *Inkerinsuomen korpus* ja *Karjalan suomen korpus*), jotka ovat jo nyt useiden tutkijoiden käytössä. Rajakarjalais- ja inkeriläismurteiden aineistot edustavat 1900-luvun jälkipuoliskolla tallennettua vanhan polven puhekieltä, kun taas Petroskoin suomen aineisto koostuu 2000-luvun mediakielestä.

Näitä korpuksia yhdistää se, että niiden edustamat kielimuodot ovat olleet omalla alueellaan vähemmistökieliä ja niihin on tullut runsaasti kontaktivaikutusta seudun valtakielestä. Inkerinsuomi ja Petroskoin (yleiskielinen) suomi ovat siirtolaiskielimuotoja, jotka ovat syntyneet siirtolaisryhmien muuttaessa uudelle kielialueelle: inkerinsuomalaisten esi-isät siirtyivät 1600-luvulla Kannakselta ja Savosta Inkerinmaalle ja Petroskoin suomalaiset 1930-luvulla Suomesta ja Yhdysvalloista Neuvosto-Karjalaan. Molemmissa suomen muodoissa venäjän kielen vaikutus on selvää. Rajakarjalaismurteita taas on alun perin puhuttu Suomen ja Venäjän välisellä raja-alueella, jossa karjalan kieli on saanut vaikutusta sekä venäjästä että suomesta.

***Multilingual Finnic**. Language contact and change*. 425–438.
Uralica Helsingiensia 14. Helsinki 2019.
‹https://doi.org/10.33341/uh.85045›

Inkerinsuomen korpus ja Raja-Karjalan korpus ovat puolestaan siinä suhteessa samanlaisia, että molemmissa on näkyvissä kaksi kielikontaktien kerrosta. Osa kieltenvälisestä vaikutuksesta on pitkäaikaista ja peräisin tilanteesta, jossa eri kielten puhujia on asunut samoilla seuduilla, Inkerinmaalla ja Raja-Karjalan alueella. Osa aineistoissa ilmenevästä vaikutuksesta on kuitenkin tuoretta ja idiolektikohtaista ja on syntynyt haastateltujen ihmisten henkilökohtaisten kokemusten kautta heidän asetuttuaan asumaan läheistä sukukieltä puhuvaan uuteen kieliyhteisöön. Inkerinsuomalaisista informanteista osa oli haastattelutilanteessa asunut noin 50 vuotta Virossa. Rajakarjalaiset oli jatkosodan jälkeen asutettu muualle Suomeen suomenkielisille alueille, ja haastattelujen aikaan he olivat olleet Suomessa noin 20–30 vuotta.

Korpustyön tavoitteena on yhtäältä parantaa aiemmin, jo 1960-luvulta lähtien koottujen puhekielisten aineistojen käytettävyyttä ja toisaalta tarjota Karjalan suomen lehtiteksteistä yhtenäinen tutkimusaineisto, jollaista ei ennestään ole olemassa. Korpusten laadinta hyödyttää myös kieliteknologista tutkimusta: korpustyössä voidaan testata kieliteknologisten työkalujen soveltuvuutta sekakieliseen ja murteelliseen puhekielen aineistoon. Tavoitteena on, että korpukset tarjoaisivat monipuolisia mahdollisuuksia sekä perinteisiin että uudentyyppisiin murteita ja kontaktivarieteetteja koskeviin tutkimusaiheisiin. Suomalaisessa dialektologiassa on toistaiseksi ollut niukasti käytettävissä varsinaisia korpuksia, jotka mahdollistaisivat esimerkiksi yhtenäisistä teksteistä tehtävät automaattiset hakutoiminnot. Ensimmäinen korpuksen muotoon koostettu murreaineisto on *Lauseopin arkiston* kokoelma. Se käsittää noin 130 tuntia Kotimaisten kielten keskuksen Suomen kielen nauhoitearkiston litteroituja murrehaastatteluja. Kotimaisten kielten keskuksessa on lisäksi saatu valmiiksi *Suomen kielen näytteitä* -sarjan litteroitujen tekstien (100 t) yhdistäminen ääninäytteisiin. Itä-Suomen yliopiston Raja-Karjalan korpuksen ja Inkerinsuomen korpuksen lopullinen tavoite on vastaavanlainen: puheäänen kohdistaminen tarvittavalla tarkkuudella litteraatiotekstiin.

2. Raja-Karjalan korpus

Raja-Karjalan korpus perustuu Kotimaisten kielten keskuksessa säilytteillä oleviin suomen kielen nauhoitearkiston murreäänitteisiin, jotka on tallennettu pääosin 1960-luvulla ja digitoitu vuosina 2009–2011. Äänitteet edustavat luovutetun Raja-Karjalan pitäjien murteita, joiden tutkimus on tähän saakka ollut vähäistä: 1800-luvun lopulla ja 1900-luvulla on ilmestynyt yksittäisiä, lähinnä äänneopillisia kuvauksia (Genetz 1870; Kujola 1910; Turunen 1965, 1973, 1982), joiden perusteella saa yleiskuvan siitä, millaisia murre-eroja alueella on ollut ennen viime sotia. Rajakarjalaismurteet eivät kuitenkaan ole mukana esim. Bubrihin, Beljakovin ja Punžinan *Karjalan kielen murrekartastossa* (1997), ja yksityiskohtainen tieto idiolektien välisestä variaatiosta on puuttunut kokonaan.

Raja-Karjalan korpukseen kuuluu yhteensä noin 120 tuntia litteroitua vanhan ikäpolven (70–90-vuotiaiden) haastattelupuhetta seuraavista pitäjistä: Ilomantsi, Korpiselkä, Suistamo, Suojärvi, Impilahti ja Salmi. Murrenäytteiden valintaperusteena on pidetty sitä, että ne edustavat äidinkieleltään karjalankielisten puhujien murretta. Haastateltavat ovat siis syntyneet luovutetun Raja-Karjalan alueella, mutta heidät on asutettu siirtolaisina toisen maailmansodan jälkeen nykyisen Suomen rajojen sisäpuolelle, enimmäkseen Pohjois-Karjalaan ja Pohjois-Savoon. Erityisesti Ilomantsin, Korpiselän ja Impilahden kielenoppaiden puhekieli on saanut paljon vaikutteita suomen savolaismurteista, kun taas Suistamon, Suojärven ja Salmin murteiden puhujilla karjalan kieli on säilynyt paremmin.

Rajakarjalaismurteiden korpus luotiin Raja-Karjalan kielikontaktien tutkimusta varten. Valtaosa äänitteistä litteroitiin puolikarkeaa transkriptiota hieman karkeammalla tarkekirjoituksella vuosina 2009–2011 Itä-Suomen yliopistossa opiskelijavoimin, Karjalaisen Kulttuurin Edistämissäätiön rahoituksella. Kun Suomen Akatemian rahoittama nelivuotinen tutkimushanke FINKA (*Suomen ja karjalan rajalla: näkökulmia lähisukukieliin ja niiden murteisiin*) perustettiin 2011, aloitettiin litterointien tarkistustyö ja aineiston täydentäminen erityisesti Raja-Karjalan itäisimpien murteiden litteroinneilla. FINKA-hankkeen lisäksi korpustyötä on rahoitettu Koneen Säätiön apurahalla, joka myönnettiin vuosiksi 2013–2015 *Suomen*

itäpuolisten lähialueiden kielikorpukset (SILK) -hankkeelle, sekä Karjalaisen Kulttuurin Edistämissäätiön apurahalla v. 2016–2017 (SILK 2). Aineisto on nyt lopullisessa koossaan, ja se on kokonaan tarkistettu. Korpuksen koko on noin 850 000 sanaa. Korpusta käytetään FINKA-hankkeessa alkaneeseen, erityisesti suomen ja karjalan sekamurteiden morfologisten, morfosyntaktisten ja foneettis-fonologisten ilmiöiden tutkimukseen. Aineistosta on valmistunut kaksi väitöskirjatutkimusta (Kok 2016, Uusitupa 2017), siitä on valmisteilla viisi väitöskirjatutkimusta (Laura Arantola, Natalia Giloeva, Henna Massinen, Ilja Moshnikov ja Susanna Tavi) ja tehtynä on useita muita tutkimuksia (Palander & Riionheimo 2018, Uusitupa, Koivisto & Palander 2017, Palander, Riionheimo & Koivisto 2018, Koivisto 2018).

Korpus on nykyvaiheessaan digitaalisessa muodossa oleva tekstikokoelma, jossa jokaisen nauhoitteen litterointi on omana Word-tiedostonaan sekä Unicode-muotoisena tekstitiedostona.[1] Tekstitiedostot on nyt kohdistettu äänitiedostoihin. Ääninäytteet on yhdistetty teksteihin Praat-ohjelmaa käyttäen ns. puoliautomaattisella nimikointimenetelmällä. Korpuksen valmistuttua ääni- ja tekstitiedostot ovat käytettävissä rinnakkain, jolloin tutkija voi seurata litteroitua tekstiä ja autenttista murrepuhetta samanaikaisesti. Äänen ja tekstin yhdistäminen palvelee esimerkiksi lausepainon hyödyntämistä, sillä lausepainolla sekä äänensävyn ja -voimakkuuden vaihtelulla on merkitystä puheen syntaktisen rakenteen ja lausesemantiikan tutkimuksessa. Tekstin ja äänen yhdistävä korpus antaa mahdollisuuksia myös uudenlaisten tutkimuskysymysten kehittelyyn: miten suomen ja karjalan kohtaaminen on vaikuttanut rajakarjalaismurteiden intonaatioon tai rytmiin (esim. lyhyttä ensi tavua seuraavan toisen tavun lyhyen vokaalin kestoon; vrt. savolaiseen ns. puolipitkään vokaaliin: *talò* : *talòssa*).

1. Korpuksen metatiedot ovat META-SHARE-tietokannassa osoitteessa ‹http://meta-share.csc.fi/repository/browse/the-corpus-of-border-karelia/f2fdd49caac211e390f0005056be118eda6c88241c1440678c85b11488d58ae0/›.

3. Inkerinsuomen korpus

Inkerinsuomen korpus perustuu Joensuun yliopiston tutkimushankkeessa tehtyihin ja nykyisin Itä-Suomen yliopistossa säilytteillä oleviin murreäänitteisiin, jotka on tallennettu 1990-luvulla. Kielentallennusta tehtiin hankkeessa *Inkerinsuomalaisten kieliolot ja inkerinsuomen nykytila*, jota johtivat suomen kielen professori Ilkka Savijärvi ja venäjän kielen professori Muusa Savijärvi. Hankkeen lähtökohtana oli inkeriläismurteiden kohtalo 1990-luvun uudessa yhteiskunnallisessa tilanteessa Neuvostoliiton romahtamisen jälkeen ja vähemmistöemansipaation alkuvaiheissa (hankkeen taustoista ks. esim. Riionheimo 2007: 21–22). Nauhoiteaineistoa kerättiin Virossa kolmella paikkakunnalla (Tartto, Pärnu ja Järvamaan maakunta) yhteensä noin 60 tuntia ja Venäjällä inkerinsuomen alkuperäisellä puhuma-alueella Inkerinmaalla viidessä pitäjässä (Toksova, Keltto, Skuoritsa, Kupanitsa ja Narvusin Kurkolanniemi) yhteensä noin 125 tuntia. Aineistosta on tähän mennessä julkaistu kolme kielennäytekokoelmaa (Riionheimo & Kivisalu 1994, Savijärvi ym. 1996, Kokko ym. 2003).

Inkerin suomalaismurteista on olemassa jonkin verran aikaisempaa tutkimusta, vaikkakin se on ollut huomattavasti vähäisempää ja epäsystemaattisempaa kuin Suomen alueella puhuttujen murteiden tutkimus. Ennen 1990-lukua tutkimus on tapahtunut samoin menetelmin ja päämäärin kuin muidenkin suomen murteiden: päähuomio on ollut äännehistoriassa, ja morfologiaa on kuvattu vain vähän, syntaksia tuskin lainkaan. Tutkimuskohteena ovat olleet vanhat inkerinsuomen murteet sellaisina kuin niitä puhuttiin ennen sotia alkuperäisillä asuma-alueilla tiiviissä suomenkelisissä yhteisöissä, ja aineistoa on koottu joko 1900-luvun alussa tai sotien jälkeen iäkkäiltä kielenoppailta. (Inkerinsuomen aiemmasta tutkimuksesta tarkemmin esim. Kokko 2007: 25–27, Riionheimo 2007: 20–21.) 1900-luvun loppupuolella puhuttu inkerinsuomi on kuitenkin aivan toisenlainen kielimuoto kuin vanhat paikallismurteet. Inkerinsuomalaiset ovat Neuvostoliiton karkotusten ja toisen maailmansodan tapahtumien vuoksi eläneet pääasiassa alkuperäisen kotiseutunsa ulkopuolella, hajallaan eri alueilla ja eri maissa ja kaikkialla pienenä vähemmistönä toisenkielisen enemmistön keskuudessa. Kieleen ovat vaikuttaneet erilaiset muutosvoimat kuin perinteisiin murteisiin, esimerkiksi kielenvaihto

enemmistökieleen, monikielisyys ja äidinkielen hiipuminen. Inkerinsuomen tutkimus on tämän myötä suuntautunut kohti uudenlaisia näkökulmia, kuten kielikontaktien ja kielen attrition tutkimusta.

Joensuun yliopiston inkerinsuomen tutkimushankkeessa aineistonkeruun tavoitteena oli hankkia yleiskuva inkerinsuomen silloisesta tilanteesta ja kielimuodon moninaisesta vaihtelusta, ja siksi haastateltavina on ollut monenlaisia kielenpuhujia: sekä murteensa säilyttäneitä että sellaisia yksilöitä, joilla alkuperäinen murre on muuttunut joko suomen yleiskielen tai vieraan kielen (venäjän tai viron) vaikutuksesta. Suurin osa haastateltavista edustaa tuolloista vanhinta ikäpolvea eli ikäluokkaa, joka oli ehtinyt omaksua Inkerinmaalla suomen äidinkielekseen ennen toisen maailmansodan aikaista ja jälkeistä kansallista hajaannusta. Aineisto on ainutlaatuinen dokumentti erään Suomen valtion ulkopuolella puhutun suomen murteen kohtalosta: sosiopoliittisista syistä aiheutuneesta kielenvaihdosta ja sen seurauksista puhujien äidinkieleen (inkerinsuomeen). Koska inkerinsuomalaisten yhteiskunnallinen asema oli Neuvostoliitossa suhteellisen samanlainen sekä Inkerinmaan alueella että Virossa, aineiston kaksi osaa mahdollistavat vertailun kahden erilaisen kielikontaktitilanteen välillä: Virossa valtakielenä on ollut läheinen sukukieli viro, Inkerinmaalla taas suomesta typologisesti paljon poikkeava venäjän kieli.

Inkeriläismurteista on ilmestynyt kaikkiaan neljä väitöskirjaa (Lehto 1996, Kokko 2007, Riionheimo 2007 ja Mononen 2013), joista Kokon ja Riionheimon tutkimukset perustuvat Itä-Suomen yliopiston Inkerinsuomen korpukseen. Samasta aineistosta on lisäksi valmistunut pro gradu -tutkielmia (tuorein Surakka 2011). Helka Riionheimo on julkaissut väitöskirjansa jälkeen useita inkerinsuomen aineistoa hyödyntäviä artikkeleita sekä suomenkielisissä että englanninkielisissä julkaisuissa tai kokoomateoksissa. Uusin vaihe aineiston hyödyntämisessä on Riionheimon ja Maria Frickin yhteistyö, jossa Viron inkerinsuomalaisten aineistoa on verrattu 1990-luvun jälkeen Viroon muuttaneiden suomalaissiirtolaisten käyttämään kieleen (Frick & Riionheimo 2013, Riionheimo & Frick 2014). Tämä yhteistyö on osoittanut myös sen, että haastatteluformaatista huolimatta inkerinsuomen aineisto on kiinnostava tutkimuskohde myös vuorovaikutus(sosio)lingvistiikan näkökulmasta. Lisäksi Riionheimolta on ilmestynyt *Virittäjä*-lehdessä yhteisartikkeli (Riionheimo ym. 2014), jossa verrataan

passiivimuotojen kohtaloa inkerinsuomen ja viron kielikontaktissa siihen, mitä suomen passiiville tapahtuu kääntämisessä, ja siihen, miten suomen passiivi vaikuttaa englannin kielen passiivin omaksumiseen. Nämä julkaisut osoittavat, että inkerinsuomen aineistolla on edelleen runsaasti annettavaa tutkimukselle ja että aineistoa voidaan lähestyä uusista näkökulmista.

Inkerinsuomen digitaalisen korpuksen työstäminen on tätä kirjoitettaessa loppusuoralla. Työ aloitettiin kirjoittamalla eri aikoina tehtyjä alkuperäisiä (osin käsin kirjoitettuja) litteraatioita sähköiseen muotoon tekstinkäsittelyohjelmalla. Aineiston nauhoitteet digitoitiin jo 2000-luvun alussa Joensuun yliopiston suomen kielen aineistokokoelmien laajemman digitoinnin yhteydessä, mutta muu korpustyö tuli mahdolliseksi keväällä 2013 Koneen Säätiön rahoittaman CROSSLING-hankkeen puitteissa. CROSSLING- ja SILK-hankkeiden mahdollistamien tutkimusapulaisen työkuukausien aikana kaikki olemassa olevat litteraatiot (noin 92 tuntia) kirjoitettiin tietokonemuotoon Word-tiedostoiksi. Litteraatioita olivat tehneet useat eri litteroijat eri aikoina eri tarkoituksiin, ja puhtaaksikirjoittamisen aikana käytettyä tarkemerkistöä yksinkertaistettiin jonkin verran. Korpustyö on viimeistelty FIN-CLARINin rahoituksella kesällä 2019. Viimeisessä vaiheessa eri tarkkuusasteilla tehdyt litteraatiot on karkeistettu yhdenmukaisiksi ja muunnettu txt-tiedostoiksi. Aineistot on luovutettu Kielipankkiin[2], ja korpuksen viimeistely siellä aloitetaan syksyllä 2019. Inkerinsuomen korpus julkaistaan tekstikorpuksena, jossa myös äänitiedostot ovat tutkijoiden saatavilla. Tekstin ja äänen kohdistaminen on hidas ja suuritöinen urakka, johon ei inkerinsuomen korpuksen osalta ole ryhdytty.

2. Tiedot korpuksesta on jo liitetty FIN-CLARINin META-SHARE-tietokantaan: ‹http://meta-share.csc.fi/repository/browse/the-corpus-of-ingrian-finnish/0bed3e04aacb11e390f0005056be118e57c9201eecd4428a9e86b7ac323f8ea8/›.

4. Karjalan suomen korpus

Suomalaisia ja suomea puhuvaa väestöä on Karjalan tasavallassa ja Petroskoissa asunut koko 1900-luvun ajan. Suomen itsenäistymisen jälkeen suomalaisia muutti paljolti poliittisista syistä Neuvostoliittoon, ja suomalaisten määrä tasavallassa nousi tasaisesti vuoden 1926 väestölaskennan 2 544:stä vuoden 1959 väestölaskennan 27 829:ään (Vsesojuznaja perepis naselenija 1926 goda, Vsesojuznaja perepis naselenija 1959 goda). Sen jälkeen suomalaisten osuus on taas tasaisesti vähentynyt, niin että vuoden 2010 väestölaskennan mukaan suomalaisia on koko tasavallassa 8 577 (Vserossijskaja perepis naselenija 2010). Petroskoissa suomen kielellä ilmestyy sanomalehti *Karjalan Sanomat* ja kulttuuriaikakauslehti *Carelia*, ja kaupungissa toimii mm. suomalainen teatteri.

Karjalan suomen korpus on koottu Venäjän Karjalassa Petroskoissa ilmestyvän *Karjalan Sanomat* -sanomalehden teksteistä kahta tarkoitusta varten: Karjalan suomen ja käännetyn suomen variaation tutkimukseen. Perusteena kyseisen aineiston valinnalle on se, että Karjalan suomea ei ole toistaiseksi koottu elektroniseksi korpukseksi. Mediatekstien voidaan katsoa edustavan normia luovaa osaa kielenkäytöstä. Karjalan suomen korpuksen edustama kielimuoto on lähellä suomen yleiskieltä.

Karjalan suomen korpuksen perustana ovat digitaaliset aineistot, joiden kokoaminen aloitettiin Suomen Akatemian ja Venäjän humanistisen tiedesäätiön vuosina 2009–2011 rahoittamassa tutkimushankkeessa *Venäjästä suomeksi ja suomesta venäjäksi: kääntäminen monikulttuurisessa yhteisössä*. Hanke toteutettiin yhteistyönä Joensuun yliopiston (nykyisen Itä-Suomen yliopiston) humanistisen osaston ja Petroskoin valtiollisen yliopiston suomen kielen ja kirjallisuuden laitoksen kanssa. Yhtenä hankkeen osa-alueena oli koota kirjoitettua Venäjän Karjalan suomea elektroniseksi korpukseksi, jonka pohjalta voisi analysoida vähemmistökielen erityispiirteitä kääntämisen ja kielikontaktien näkökulmasta sekä verrata aineistoa vastaavaan Suomessa tuotettuun suomenkieliseen materiaaliin.

Karjalan suomen tekstikorpus saatiin valmiiksi Koneen Säätiön rahoittaman SILK-hankkeen aikana vuosina 2013–2014 yhteistyössä Petroskoin valtiollisen yliopiston kanssa. Korpus sisältää noin 600 000

sanaa *Karjalan Sanomat* -sanomalehden 2000-luvulla ilmestyneistä teksteistä. Kukin aineiston artikkeli on annotoitu käännöstieteellisen ja kielikontaktien tutkimuksen kannalta relevantilla tavalla, josta käy ilmi artikkelin syntytapa, toisin sanoen, onko kyseessä käännetty vai alun perin suomeksi tuotettu aineisto. Analyysissa on hyödynnetty myös venäjänkielisiä lähdetekstejä. Korpus[3] on nyt Unicode-muotoon koodattuina tekstitiedostoina, ja se mahdollistaa seuraavassa vaiheessa tekstin morfologis-syntaktisen annotoinnin. Hankkeen käyttöoikeus on ensi vaiheessa sekä Petroskoin valtiollisella yliopistolla että Itä-Suomen yliopistolla.

Jukka Mäkisalo, Hannu Kemppanen ja Anna Saikonen (2016) ovat esitelleet Karjalan suomen korpusta ja ensimmäisiä siitä tehtyjä korpusanalyyseja käännöstieteellisessä *MikaEL*-julkaisussa. Käännetyn ja ei-käännetyn kieliaineiston vertailu vähemmistökielen näkökulmasta on kyseenalaistanut aiempia väittämiä näille kielimuodoille tyypillisistä piirteistä.

Kiitokset

Korpushankkeemme ovat saaneet taloudellista tukea seuraavilta rahoittajilta: Suomen Akatemialta (137479; Raja-Karjalan korpus), Koneen Säätiöltä (40-5091; Raja-Karjalan, Inkerinsuomen ja Karjalan suomen korpukset), Karjalaisen Kulttuurin Edistämissäätiöltä (Raja-Karjalan ja Inkerinsuomen korpukset) sekä FIN-CLARINilta (Inkerinsuomen korpus). Kiitämme lämpimästi hankkeille osoitetusta tuesta.

3. Korpuksen metatiedot ovat FIN-CLARINin META-SHARE-tietokannassa osoitteessa ‹http://meta-share.csc.fi/repository/browse/the-karelian-finnish-newspaper-corpus/80fa56f0454e11e49821005056be118e6a793e3276d84c95b8d9cf6ff7d867c8/›.

Lähteet

Bubrih, D. V., A. A. Beljakov & A. V. Punžina 1997: *Karjalan kielen murrekartasto. Dialektologičeskij atlas karelskogo jazyka* [mukana tverinkarjalan murrekartat]. Toim. Leena Sarvas. Venäjän tiedeakatemian Karjalan tiedekeskuksen kielen, kirjallisuuden ja historian instituutti & Kotimaisten kielten tutkimuskeskus. Kotimaisten kielten tutkimuskeskuksen julkaisuja 97. Helsinki: Suomalais-Ugrilainen Seura.

Frick, Maria & Helka Riionheimo 2013: Bilingual voicing: A study of code-switching in the reported speech of Finnish immigrants in Estonia. – *Multilingua* 32:5: 565–599.

Genetz, Arvid 1870: Kertomus Suojärven pitäjäästä ja matkustuksistani siellä v. 1867. – *Suomi II: 8. Kirjoituksia isän-maallisista aineista*. Helsinki: SKS.

Koivisto, Vesa 2018: Border Karelian dialects – a diffuse variety of Karelian. – Marjatta Palander, Helka Riionheimo & Vesa Koivisto (eds), *On the Border of Language and Dialect*. Studia Fennica Linguistica 21. Helsinki: Finnish Literature Society. 56–84. Saatavissa: ‹http://dx.doi.org/10.21435/sflin.21›

Kok, Maria 2016: *Varjon kieliopillistuminen.* Itse-*sanan paradigman rakenne ja merkityksenkehitys itäisessä itämerensuomessa*. [Väitöskirja.] Publications of the University of Eastern Finland. Dissertations in Education, Humanities, and Theology 83. Joensuu: Itä-Suomen yliopisto. Saatavissa: ‹http://urn.fi/URN:ISBN:978-952-61-2064-5›

Kokko, Ossi 2007: *Inkerinsuomen pirstaleisuus. Eräiden sijojen kehitys murteen yksilöllistymisen kuvastajana*. Joensuun yliopiston humanistisia julkaisuja 48. Joensuu: Joensuun yliopisto. Saatavissa: ‹http://urn.fi/URN:ISBN:978-952-219-036-9›

Kokko, Ossi, Ilkka Savijärvi & Muusa Savijärvi (toim.) 2003: *Ennev vanhasii – Pohjois-Inkerin kieltä ja kohtaloita*. Studia Carelica Humanistica 18. Joensuu: Joensuun yliopisto.

Kujola, Joh. 1910: *Äänneopillinen tutkimus Salmin murteesta*. Eripainos Suomi-kirjasta. Helsinki: SKS.

Lehto, Manja Irmeli 1996: *Ingrian Finnish: Dialect preservation and change*. [Väitöskirja.] Acta Universaliensis Upsaliensis. Studia Uralica Upsaliensia 23. Uppsala: Uppsala University.

Mononen, Kaarina 2013: *Inkerinsuomalaisten suomen kielen käyttö Pietarissa ja sen lähialueella*. [Väitöskirja.] Helsinki: Helsingin yliopiston suomen kielen, suomalais-ugrilaisten ja pohjois-

maisten kielten ja kirjallisuuksien laitos. Saatavissa: ‹http://urn.fi/URN:ISBN:978-952-10-8657-1›

Mäkisalo, Jukka, Hannu Kemppanen & Anna Saikonen 2016: *Karjalan Sanomat* -korpus. Petroskoin (käännös)suomen piirteitä. *MikaEL*. Kääntämisen ja tulkkauksen tutkimuksen symposiumin verkkojulkaisu, vol. 9. Saatavissa: ‹http://www.sktl.fi/liitto/seminaarit/mikael-verkkojulkaisu/›

Palander, Marjatta & Helka Riionheimo 2018: How is Karelian recalled and imitated by Finns with Border Karelian roots? – Marjatta Palander, Helka Riionheimo & Vesa Koivisto (eds), *On the Border of Language and Dialect*. Studia Fennica Linguistica 21. Helsinki: Finnish Literature Society. 85–122. Saatavissa: ‹http://dx.doi.org/10.21435/sflin.21›

Palander, Marjatta, Helka Riionheimo & Vesa Koivisto (eds) 2018: *On the Border of Language and Dialect*. Studia Fennica Linguistica 21. Helsinki: Finnish Literature Society. Saatavissa: ‹http://dx.doi.org/10.21435/sflin.21›

Riionheimo, Helka 2007: *Muutoksen monet juuret. Oman ja vieraan risteytyminen Viron inkerinsuomalaisten imperfektinmuodostuksessa*. [Väitöskirja.] Suomalaisen Kirjallisuuden Seuran toimituksia 1107. Helsinki: SKS.

Riionheimo, Helka & Maria Frick 2014: Emergence of Finnish-Estonian bilingual constructions in two contact settings. – *Sociolinguistic Studies* 8:3: 409–447.

Riionheimo, Helka & Krista Kivisalu (toim.) 1994: *Inkeriläiskertomuksia*. Studia Carelica Humanistica 4. Joensuu: Joensuun yliopisto.

Riionheimo, Helka, Leena Kolehmainen & Lea Meriläinen 2014: Suomen passiivi kontaktissa. Kieltenvälisiä kytköksiä migraatiossa, toisen kielen omaksumisessa ja kääntämisessä. – *Virittäjä* 118: 334–371. Saatavissa: ‹https://journal.fi/virittaja/article/view/9249›

Savijärvi, Ilkka, Muusa Savijärvi & Janne Heikkinen (toim.) 1996: *Vot, ihminen tahtoo kotimaalle. Länsi-Inkerin kieltä ja kohtaloita*. Studia Carelica Humanistica 8. Joensuu: Joensuun yliopisto.

Surakka, Anne 2011: *Yleistävän yksikön 2. persoonan käyttö inkerinsuomessa*. Pro gradu -tutkielma. Itä-Suomen yliopisto, suomen kieli. Saatavissa: ‹http://urn.fi/urn:nbn:fi:uef-20110436›

Turunen, Aimo 1965: Suojärven murre. – Lauri Pelkonen (toim.), *Suojärvi I*. Kajaani: Suo-säätiö. 21–38.

—— 1973: Raja-Karjalan murteet ja vepsän kieli. – Hannes Sihvo (toim.), *Kalevalaseuran vuosikirja 53*. Helsinki: SKS. 83–94.

—— 1982: Raja-Karjalan murteet. – *Karjala 2. Karjalan maisema ja luonto*. Hämeenlinna: Karisto. 65–89.

Uusitupa, Milla 2017: *Rajakarjalaismurteiden avoimet persoonaviittaukset*. [Väitöskirja.] Publications of the University of Eastern Finland. Dissertations in Education, Humanities, and Theology 117. Joensuu: Itä-Suomen yliopisto. Saatavissa: ‹http://urn.fi/URN:ISBN:978-952-61-2646-3›

Uusitupa, Milla, Vesa Koivisto & Marjatta Palander 2017: Raja-Karjalan murteet ja raja-alueiden kielimuotojen nimitykset. – *Virittäjä* 121: 67–106. Saatavissa: ‹https://journal.fi/virittaja/article/view/53121›

Vserossijskaja perepis naselenija 2010. Natsionalnyi sostav naselenija po subjektam Rossijskoi federatsii [Koko Venäjän kattava väestölaskenta 2010. Osa 1. Väestön lukumäärä ja jakauma. Taulukko 7 [MS Excel-taulukko]]. Federalnaja služba gosudarstvennoi statistiki [Venäjän federaation tilastovirasto], 2012. Moskova: ИИЦ ”Статистика России”. [Viitattu 18.12.2015] Saatavissa: ‹http://www.gks.ru/free_doc/new_site/population/demo/per-itog/tab7.xls›

Vsesojuznaja perepis naselenija 1926 goda. Natsionalnyi sostav naselenija po regionam RSSR [Koko Neuvosto-Venäjän kattava väestölaskenta 1926. Väestön kansallisuus ja jakautuminen maaseutu- tai kaupunkiasukkaisiin. Karjalan ASSR]. Демоскоп Weekly. [Viitattu 18.12.2015] Saatavissa: ‹http://demoscope.ru/weekly/ssp/rus_nac_26.php?reg=53›

—— 1959 goda. Natsionalnyi sostav naselenija po regionam Rossii [Koko Neuvostoliiton kattava väestölaskenta 1959. Väestön kansallisuus Venäjän alueilla. Karjalan ASSR]. Демоскоп Weekly. [Viitattu 18.12.2015] Saatavissa: ‹http://demoscope.ru/weekly/ssp/rus_nac_59.php?reg=81›

Language corpora from the eastern border of Finland

Marjatta Palander, Helka Riionheimo, Hannu Kemppanen & Jukka Mäkisalo

This report presents three language corpora on language varieties used on both sides of the eastern border of Finland: the Corpus of Border Karelia, the Corpus of Ingrian Finnish, and the Karelian Finnish Newspaper Corpus. These digital corpora have been (and are in the process of being) compiled on the subjects of the Finnish language, Russian language, and Karelian language and culture at the University of Eastern Finland. The varieties in question have been small minority languages in their respective areas, and thus, the corpora offer perspectives into the fates of minority languages and cross-linguistic influence from the dominant languages.

The Corpus of Border Karelia consists of 120 hours of transcribed dialect samples from Karelian speakers who were born in the Border Karelia parishes of Ilomantsi, Korpiselkä, Suistamo, Suojärvi, Impilahti, and Salmi. The recordings were conducted mainly in the 1960s in Finland in the places where the inhabitants of Border Karelia were resettled after World War II (i.e., when Finland ceded the Border Karelia region to the Soviet Union). The samples were transcribed in 2009–2014 and the Unicode texts and sound files are currently being aligned.

The interviews that form the basis of the Corpus of Ingrian Finnish were recorded in the 1990s in two locations: in Estonia and in Russia (in the area named Ingria, surrounding St. Petersburg). The corpus consists of 125 hours of recordings from Ingria (the parishes of Toksova, Keltto, Skuoritsa, Kupanitsa, and Kurkolanniemi) and about 60 hours of recordings made in Estonia (in the towns of Tartu and Pärnu and the Järvamaa district). Parts of this data have been transcribed during the last two decades by several transcribers. At present, we are in the process of transforming these miscellaneous (partly hand-written) texts into digital form in order to build them into a text corpus in the Unicode format. In the future, it will be possible to align the text and sound files as well.

The Karelian Finnish Newspaper Corpus is a text corpus which comprises written texts that have been published in the *Karjalan Sanomat* newspaper in Petrozavodsk (in the Karelian Republic of the Russian Federation) in the 2000s. The corpus was compiled in 2009–2014 in cooperation with the State University of Petrozavodsk. It contains about 600 000 words, and all the texts have been annotated so that we know whether the text was translated from Russian or originally written in Finnish. Work on this corpus has now been completed and it has been handed over to the FinClarin database.

Ultimately, the Corpus of Border Karelia and the Corpus of Ingrian Finnish will also be included in FinClarin. The metadata of these corpora have already been included in the META-SHARE database.