ANTTI LEINO
SAARA HYVÖNEN
MARKO SALMENKIVI

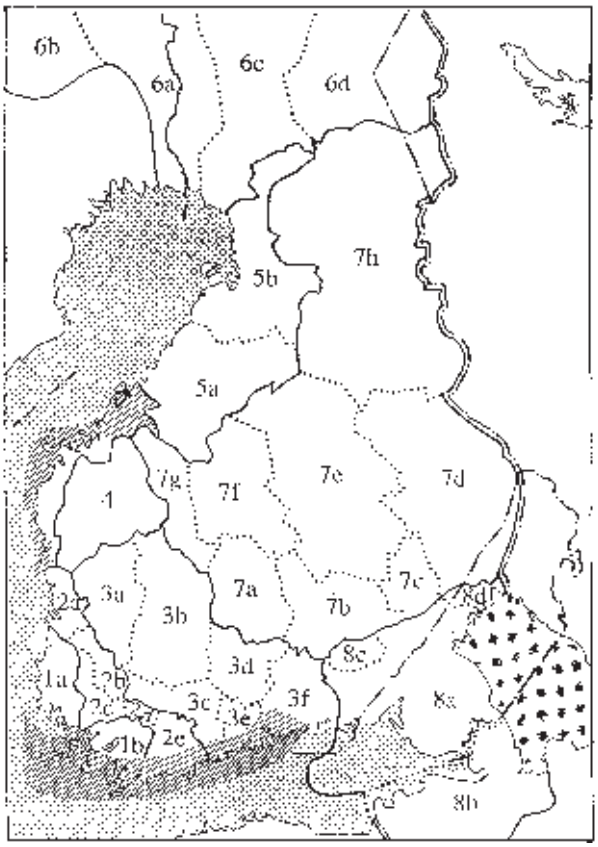
MITÄ MURTEITA SUOMESSA ONKAAN?

MURRESANASTON LEVIKIN KVANTITATIIVISTA ANALYYSIÄ

Käsitys suomen murteista on säilynyt kohtuullisen muuttumattomana puolisen vuosisataa. Vakiintunut kuva perustuu olennaisesti siihen äänne- ja muotopiirteiden joukkoon, joka on kuvattuna Lauri Kettusen (1940) murrekartastossa. Tämä kuvassa 1 näkyvä murrejako on vuosien varrella muuttunut hyvin vähän.

Vakiintuneen käsityksen mukaan suomen murteet jakaantuvat ensin kahtia itä- ja länsimurteisiin. Kun murrejakoa tarkennetaan, päästään kahdeksaan murrealueeseen. Näistä kaksi — savolais- ja kaakkoismurteet — kuuluvat itämurteisiin, länsimurteita puolestaan ovat loput kuusi: hämäläismurteet, lounaismurteet, näiden väliset lounaiset välimurteet, Etelä-Pohjanmaan murteet, Keski- ja Pohjois-Pohjanmaan murteet sekä Peräpohjolan murteet.

Aivan kaikki eivät lue kaikkia kahdeksaa omiksi päämurteikseen: niinpä Rapola (1961) ei vielä mainitse lounaisia välimurteita omana ryhmänään vaan luettelee alueen toisaalta hämäläis- ja toisaalta lounaismurteiden yhteydessä; Mielikäinen (1991) puolestaan piirtää murrekarttaansa yhtenäisen pohjalaismurteiden alueen, joka sisältää Keski- ja Pohjois-Pohjanmaan lisäksi myös Etelä-Pohjanmaan murteet. Jopa itä–länsi-jaosta on käyty keskustelua: Paunonen (1991) esittää, että Keski- ja Pohjois-Pohjanmaan sekä Peräpohjolan murteet muodostavat itä- ja länsimurteiden kanssa rinnasteisen pohjoisen alueen.



Kuva 1. Suomen murrealueet (Savijärvi ja Yli-Luukko 1994).

1. Lounaismurteet
a) pohjoisryhmä, b) itäryhmä
2. Lounaiset välimurteet
a) Porin seudun, b) Ala-Satakunnan, c) Turun ylämaan, d) Someron, e) Länsi-Uudenmaan murteet
3. Hämäläismurteet
a) Ylä-Satakunnan murteet, b) perihämäläiset murteet, c) etelähämäläiset murteet, d–f) kaakkoishämäläiset murteet (d = Hollolan ryhmä, e = Porvoon ryhmä, f = Kymenlaakson eli Iitin ryhmä)
4. Etelä-Pohjanmaan murteet
5. Keski- ja Pohjois-Pohjanmaan murteet
a) Keski-Pohjanmaan, b) Pohjois-Pohjanmaan murteet
6. Peräpohjolan murteet
a) Tornion, b) Jällivaaran, c) Kemin, d) Kemijärven murteet
7. Savolaismurteet
a) Päijät-Hämeen murteet, b) Etelä-Savon murteet, c) Savonlinnan seudun välimurteet, d) Pohjois-Karjalan murteet (itäiset savolaismurteet), e) Pohjois-Savon murteet, f) Keski-Suomen murteet, g) Keuruun–Evijärven seudun välimurteet, h) Kainuun murteet
8. Kaakkoismurteet
a) varsinaiset kaakkoismurteet, b) Inkerin suomalaismurteet, c) Lappeenrannan seudun välimurteet, d) Sortavalan seudun välimurteet

▷

MILLÄ PERUSTEILLA MURTEET MÄÄRITELLÄÄN?

Vakiintunut murrejako perustuu lähinnä äänne- ja muoto-opillisten piirteiden levikkeihin. Tässä on kuitenkin omat ongelmansa: vaikka oletettaisiinkin, että kunkin piirteen levikki olisi selvärajainen, eri piirteiden isoglossit eivät useinkaan osu yhteen sillä tavoin, että murreraja voitaisiin ilman muuta määritellä. Käytännössä murrejako perustuukin »merkittävien» piirteiden levikkeihin, joita vielä painotetaan sopivasti. Tässä merkittäviksi valitaan sellaiset piirteet, että niiden perusteella syntyvä murrejako vastaa ennakko-odotuksia. Tämä menetelmällinen yksityiskohta jää usein tiedostamatta, vaikka jo Ruoppila (1937) sen jokseenkin suorasanaisesti esittää.

Myös Rapola (1961: 15) mainitsee aivan suoraan, että murreraja on »aina enemmän tai vähemmän ehdonvallan kysymys». Se, mihin raja todellisuudessa on vedettävä, on hänen mukaansa määritettävissä selvittämällä kunkin ilmiön alkuperä, ikä ja leviämishistoria — siis selittämällä murrejako siihen johtaneen historiallisen kehityksen kautta. Tällaisella maantieteellis-historiallisella menetelmällä on toki vahvuutensa, ja Tuomi (1989) toteaa aivan oikein sen osoittautuneen menestykselliseksi. Toisaalta on myös huomattava, että tällainen ennakkoaavistus »oikeasta» jaosta ja sitä kautta merkittävistä murrepiirteistä sopii jo lähtökohdiltaan kovin huonosti synkroniseen tutkimukseen: Paunonen (1991: 92–93) muistuttaakin, että lähtökohdaksi olisi tällöin otettava yksittäisten piirteiden sijaan systemaattisempi murteiden typologinen tarkastelu. Jos tavoitteena on deskriptiivinen esitys kielen alueellisesta vaihtelusta, jako on tehtävä havaitun vaihtelun pohjalta.

Murrejaon selvittäminen synkronisesti, todella havaitun vaihtelun perusteella, edellyttää luonnollisesti keinoja vaihtelun jyrkkyyden mittaamiseksi. Tietotekniikan kehitys parin viime vuosikymmenen aikana onkin antanut uusia välineitä myös murteentutkimuksen käyttöön. Suomessa tällaista dialektometrista tutkimusta on ainakin jo Palanderilla (1996), mutta tuolloin tutkimus kohdistui vielä suppeahkoon alueeseen ja vakiintuneen murrejaon kannalta olennaisiin piirteisiin. Varsinaista koko murteiston yleiskatsausta on yrittänyt näkyvimmin Wiik (2004). Hänen lähtökohtansa on sikäli aidon dialektometrinen, että tutkimuksessa on tarkasteltu kaikkia Kettusen (1940) esittämiä murrepiirteitä ja mitattu murrerajojen jyrkkyyttä laskemalla yhteen lankeavien isoglossien lukumäärät. Toisaalta kaikki Kettusen kartat on otettu mukaan samanarvoisina, vaikka joukkoon mahtuu kahden tai useammankin kartan sarjoja, jotka kuvaavat olennaisesti saman ilmiön levikkiä, kuten Palander (1999: 263) on jo teoksen käsikirjoituksesta huomauttanut; tällaisia kartaston päällekkäisyyksiä käsittelee tarkemmin Mielikäinen (1990). Lisäksi Wiikin tutkimuksessa on edelleen taustalla vakiintunut tieto murrejaosta, mikä näkyy toisaalta kirjan jäsentelyssä, toisaalta eri murrealueiden vaikutuksen voimakkuutta kuvaavissa kartoissa. Kettusen murrekartastoa ovat tutkineet kvantitatiivisesti myös Embleton ja Wheeler (1997, 2000), mutta tämän tutkimuksen tuloksia on esitelty valitettavan vähän.

Tässä työssä on menty vielä askelta pidemmälle sikäli, että itse analyysiä tehtäessä ei ole käytetty minkäänlaista tietoa aiemmista murrejaoista tai edes siitä, missä päin Suomea kukin pitäjä sijaitsee. Tarkoituksena on ollut selvittää, löytyykö suuresta joukosta sanojen levikkejä sellaisia säännönmukaisuuksia, jotka voidaan tulkita murrerajoiksi. Puhtaan aineistolähtöisyyden lisäksi uutta tutkimuksessamme on myös itse aineisto: aiempi murteentutkimushan on Suomessa keskittynyt valtaosaltaan äänne- ja muoto-opillisiin piirteisiin, osittain siksi, että niiden levikkitietoa on ollut saatavilla, osittain siksi,

että tällaisten vaihteluiden selvittäminen on ylipäätään nähty murteentutkimuksen keskeiseksi tehtäväksi, ja osittain fennistiikassa vallinneen vahvan kieli- ja erityisesti äännehistoriallisen pohjavireen johdattamana.

Toisaalta on pidettävä mielessä, että sanastoilmiöt leviävät kieliyhteisössä eri tavoin kuin äänne- ja muoto-opilliset piirteet, kuten muiden muassa Savijärvi (1995) toteaa Jämsän seudun murteista ja Haspelmath (2004: 212) puolestaan hiukan nyt käsillä olevaa murretutkimusta yleisemmin. Suomen murteistossa tämä näkyy selvästi eri siirtymämurteissa, joiden sanasto viittaa useinkin vastakkaiseen suuntaan kuin äänne- ja muoto-opilliset piirteet. Ilmiö nousee esiin myös omassa tutkimuksessamme, eikä ole varsinaisesti kovin yllättävää, että sanastoon perustuva murrejakomme eroaa äänne- ja muotopiirteisiin perustuvasta paljolti juuri eri siirtymämurteiden kohdalla.

Jo Rapola (1961: 30–31) arveli toiveikkaana, että tuolloin vauhdikkaasti edennyt murresanaston keruu mahdollistaisi ennen pitkää sanaston käytön murteiden tutkimuksessa ja että myöhemmin myös syntaktista vaihtelua voitaisiin selvittää. Jälkimmäiseen aineistomme ei anna mahdollisuuksia, mutta edelliseen olemme viimein Sanakirjasäätiön lakkauttamisen aikoihin pääsemässä käsiksi.

MENETELMÄT

Aineistomme koostuu Suomen murteiden sanakirjan toimitustyön yhteydessä piirretyistä levikkikartoista. Käytössämme on ollut runsaan 5 600 sana-artikkelin materiaali, josta — kun otetaan huomioon sanojen eri merkitysten erilaiset levikit — saadaan yhteensä noin 9 000 levikkikarttaa. Aineistoa on kerätty noin 1900-luvun alusta alkaen; keruun alkuvaiheita on tarkemmin selvittänyt Strandberg (2004).

Taulukossa 1 on esitettynä pieni otos aineistostamme. Siinä levikit on esitetty taulukkona, jonka rivit ovat sanoja tai niiden eri merkityksiä ja sarakkeet pitäjiä. Kussakin taulukon alkiossa on arvo 1, jos sana on kerätty pitäjästä, ja 0, jos näin ei ole. Tätä aineistoa olemme tutkineet tietojenkäsittelytieteellisen data-analyysin menetelmin, joihin kohtalaisen kattava johdatus on esimerkiksi Hand ym. (2001).

	Vihti	**Aura**	**Kitee**	**Juva**
aprakka	0	0	1	1
epatto	0	0	1	1
filunki	1	1	0	0
haalakka	0	0	1	1
hampuusi	1	1	0	0
kräki	0	1	0	0

Taulukko 1. Pieni otos aineistosta. Taulukossa arvo 1 kertoo sanan esiintyvän kyseisessä kunnassa.

Suhde aiempaan dialektometriaan

Tutkimuksemme on lähtökohdiltaan läheistä sukua dialektometrialle, ainakin jos tämä termi käsitetään laajasti laskennalliseksi murteentutkimukseksi. Tällainen laskennallinen — tai oikeastaan tilastollinen — lähestymistapa ei sinänsä ole uusi: esimerkiksi Palander (1999) mainitsee varhaisia kvantitatiivisia murretutkimuksia jo 1900-luvun alusta.

Johtavana ajatuksena tähänastisessa dialektometriassa on, että murteentutkimuksen keskeiseksi menetelmäksi otettaisiin murteiden välisten erojen laskeminen. Olennaisia uudistuksia tässä on kaksi: Ensiksikin eroja tarkastellaan laskennallisesti ja ainakin periaatteessa ilman, että ennakkoaavistusten perusteella keskityttäisiin vain muutamaan »tärkeään» piirteeseen. Toiseksi tarkastelun perusyksiköksi otetaan piirteen sijasta paikallismurre, eli jos levikit esitetään taulukkona, jonka akseleina ovat toisaalta piirteet ja toisaalta paikkakunnat, taulukon akselit vaihdetaan keskenään.

Murteiden välisen etäisyyden laskemiseen dialektometrisessa tutkimuksessa on käytetty erilaisia tapoja. Viime aikoina suosituksi on tullut menetelmä, jossa eri murteenpuhujilta nauhoitetaan sama tekstikatkelma. Murteiden välinen ero saadaan tästä laskemalla niin sanottu Levenštein-etäisyys (Levenshtein 1966), siis niiden yhden (transkriptiossa käytetyn) merkin mittaisten muutosten määrä, joiden avulla yhden murteenpuhujan näyte saadaan identtiseksi toisen puhujan näytteen kanssa.

Levenštein-etäisyyteen perustuvaa dialektometrista menetelmää on sovellettu suhteellisen menestyksellisesti ainakin hollannin (Nerbonne ja Heeringa 2001; Nerbonne 2003) ja norjan (Gooskens ja Heeringa 2004) murteisiin. Omaan tutkimukseemme tällainen etäisyysmitta ei luonnollisesti sovellu, vaan etäisyydet on määriteltävä eri ilmiöiden — olivat ne sitten murrepiirteitä tai sanoja — levikkien perusteella. Hiukan samansuuntaista lähestymistapaa ovat suomen murteisiin soveltaneet ainakin Palander, Opas-Hänninen ja Tweedie (2003) sekä Wiik (2004). Omassa tutkimuksessamme emme kuitenkaan tyydy laskemaan pitäjien (tai pitäjämurteiden) välille vain yhtä etäisyyttä vaan pyrimme etsimään murteistosta erilaisia vaihteluita.

Ryvästys

Murrealueiden määrittelyyn käytämme ryvästykseksi kutsuttua menetelmää. Siinä on tarkoituksena jakaa aineisto ryppäisiin siten, että samankaltaiset oliot — tässä tapauksessa pitäjät — päätyvät samaan ryppääseen, kun taas eri ryppäisiin päätyvät oliot ovat keskenään erilaisia. Samankaltaisuus tarkoittaa tässä tietenkin sanastollista yhdenmukaisuutta. Esimerkiksi taulukon 1 aineistoa tarkastelemalla on helppo havaita, että itäsuomalaisten Juvan ja Kiteen sanajakaumat ovat identtiset ja eroavat jyrkästi länsisuomalaisten Auran ja Vihdin sanajakaumista, jotka taas muistuttavat toisiaan. Kahden rypään jako siis tuottaisi itäisen ja läntisen rypään. Jako kolmeen taas erottaisi Auran ja Vihdin omiksi rypäikseen.

Ryvästyksen kannalta aineistossamme on kaksi erityispiirrettä, jotka estävät standardimenetelmien soveltamisen menestyksekkäästi. Ensinnäkin aineisto on suuri, minkä vuoksi menetelmät toimivat kovin hitaasti, jos ollenkaan. Toiseksi sanoja on kerätty hyvin epätasaisesti eri pitäjistä, mikä helposti johtaa siihen, että vähäsanaiset pitäjät päätyvät samaan vähäsanaisten pitäjien rypääseen, mikä taas ei ole mielekästä: emme ole kiinnostuneita erottelemaan pitäjiä sen perusteella, paljonko niistä on kerätty sanoja. Sanastusaste

on kuitenkin varsin tarkasti eristettävissä muusta vaihtelusta, joten se voidaan jättää ryvästyksessä huomiotta.

Ongelmamme muistuttaa tekstidokumenttien analysointia (esim. Berry ym. 1999). Näissä sovelluksissa niin sanoja kuin dokumentteja voi olla tuhansia, ja dokumentit halutaan luokitella sanaston mukaan samaa aihepiiriä edustavien dokumenttien rypäiksi. Dokumentit voivat olla keskenään hyvin eripituisia ja näin ollen sisältää hyvin eri määrän sanoja. Tekstidokumenttien analysointiin on kehitetty erilaisia tapoja esikäsitellä aineisto. Olemme soveltaneet aineistoon näitä esikäsittelytapoja hieman muunnettuna (Hyvönen ym. tulossa). Tähän muokattuun aineistoon voidaan sitten soveltaa ryvästysmenetelmiä, tavallisimmin niin sanottua K-means-klusterointia (MacQueen 1967). Tekstidokumenttien analysoimisessa käytetään yleensä niin sanottua kosinietäisyyttä yleisemmän euklidisen etäisyyden sijaan, kun mitataan dokumenttien samankaltaisuutta. Toinen yleinen ryvästysmenetelmä, hierarkkinen klusterointi, ei tunnu tähän aineistoon sopivan, sillä hyvin erilaiset kunnat poimiutuivat tavallisesti omiksi yhden tai muutaman kunnan rypäikseen.

Ryvästyksessä jaamme pitäjät ennalta määrättyyn lukumäärään rypäitä siten, että saman rypään pitäjät ovat sanastoltaan samankaltaisia, kun taas eri rypäissä olevat pitäjät ovat sanastoltaan erilaisia. Kun rypäiden lukumäärä on pieni, saadaan pitäjille suhteellisen karkea jako: esimerkiksi kahteen rypääseen jaettaessa saadaan perinteinen jako itä- ja länsimurteisiin. Kun rypäiden lukumäärää kasvatetaan, saadaan kartalle näkyviin yhä hienojakoisempia murrealueiden osituksia.

Kullekin rypäälle voidaan siihen kuuluvien pitäjien sanastojen perusteella muodostaa keskimääräinen sanajakauma. Kunkin rypään pitäjien sanastot ovat lähempänä oman rypäänsä keskimääräistä sanajakaumaa kuin minkään vieraan rypään; tämä itse asiassa määrittää rypään. On hyvä huomata, että tämä ei tarkoita sitä, etteivätkö kahteen eri rypääseen kuuluvat pitäjät voisi käytännössä olla hyvinkin samankaltaisia — rypäiden väliset rajat eivät ole välttämättä kovin jyrkkiä, mikä näkyy selvimmin, kun katsoo, miten rypäiden lukumäärän lisääminen muuttaa murrekarttaa.

Toinen tärkeä huomio on, että menetelmät eivät käytä hyväkseen tietoa pitäjän sijainnista: maantieteellisesti yhtenäiset rypäät saadaan siksi, että murrealueet, jotka ovat sanastoltaan yhtenäisiä, ovat myös maantieteellisesti yhtenäisiä, ei siksi, että menetelmä ottaisi pitäjien maantieteellistä etäisyyttä millään tavalla huomioon.

Pääkomponenttianalyysi

Aineistossamme on tuhansia sanoja ja satoja pitäjiä, joten koko aineiston havainnollistaminen on hankalaa, mutta toisaalta jako murrealueisiin esittää vaihtelun ehkä jo tarpeettomankin paljon yksinkertaistettuna. Ratkaisuksi tähän ongelmaan olemme valinneet pääkomponenttianalyysiksi kutsutun menetelmän. Siinä tarkoitus on muodostaa pieni määrä uusia, keinotekoisia muuttujia, jotka säilyttävät mahdollisimman suuren osan aineiston vaihtelusta. Näitä keinotekoisia muuttujia sanotaan pääkomponenteiksi, ja niitä voi käyttää muun muassa aineiston havainnollistamiseen. Menetelmä on tunnettu jo pitkään — perusajatuksen merkittävimmän vaihtelusuunnan löytämisestä esitti Pearson (1901) ja analyysimenetelmän olennaisesti nykymuodossaan Hotelling (1933) — mutta tietotekniikan kehittyminen viime vuosikymmeninä on nostanut sen yhdeksi tietojenkäsittelytieteellisen data-analyysin perustekniikoista.

▷

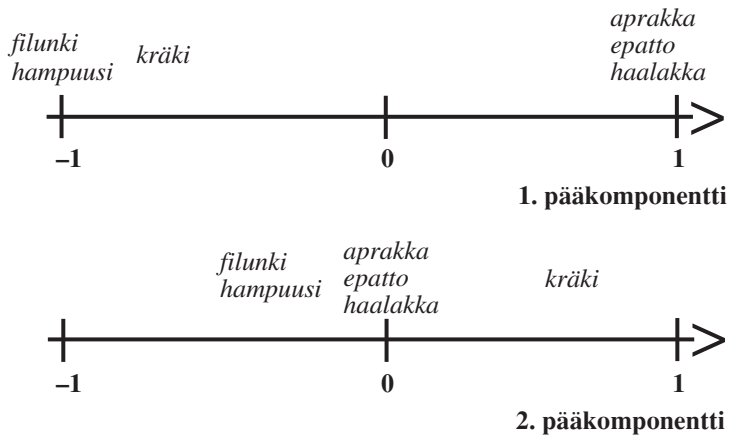
Kukin pääkomponentti on sellainen yhdistelmä kaikista alkuperäisen aineiston muuttujista, jossa kukin muuttuja on mukana omalla painokertoimellaan. Nämä painotukset on laskettu niin, että vaihtelu saadaan mahdollisimman suureksi. Alkuperäiset muuttujat ovat tässä tapauksessa pitäjiä, ja etsimme siis sellaisia pitäjien yhdistelyjä, jotka selittävät mahdollisimman suuren osan sanaston vaihtelusta. Pohjimmiltaan tämä on varsin lähellä perinteisen murteentutkimuksen menetelmiä: murrerajoja piirrettäessä eri piirteitä painotetaan niin, että lopputuloksena syntyvät rajat sopivat kieli- ja asutushistoriallisiin tietoihin; tässä sen sijaan kutakin komponenttia määritettäessä painotukset lasketaan niin, että vaihtelu on mahdollisimman suurta.

Taulukon 1 esimerkkiaineistolla saadut pääkomponentit on esitetty taulukossa 2. Jokainen pääkomponentin alkio kertoo, millä painolla sitä vastaava pitäjä on mukana kyseisessä pääkomponentissa. Itse pääkomponentti taas pyrkii selittämään sanaston vaihtelua. Ensimmäinen pääkomponentti, joka selittää 75 prosenttia aineiston vaihtelusta, erottaa länsisuomalaiset pitäjät (negatiiviset arvot) itäsuomalaisista (positiiviset arvot). Toinen pääkomponentti, joka selittää loput 25 prosenttia, erottaa Vihdin, josta on vähemmän sanoja, muusta aineistosta, erityisesti toisesta länsisuomalaisesta pitäjästä Aurasta.

Pitäjä	**1. pääkomponentti**	**2. pääkomponentti**
Vihti	–0,40	–0,91
Aura	–0,53	0,23
Kitee	0,53	–0,23
Juva	0,53	–0,23

Taulukko 2. Taulukon 1 aineistolla saadut pääkomponentit.

Ensimmäinen pääkomponentti on siis sellainen yhdistelmä alkuperäisiä muuttujia, tässä siis pitäjiä, joka säilyttää mahdollisimman suuren osan aineiston vaihtelusta. Esimerkkitapauksessamme tätä havainnollistaa kuva 2. Pääosa aineiston vaihtelusta selittyy itäisen ja läntisen sanaston eroilla, ja niinpä itäiset ja läntiset sanat sijoittuvat ensimmäisellä pääkomponentilla kauas toisistaan. Toinen pääkomponentti taas pyrkii selittämään mahdollisimman suuren osan jäljelle jääneestä vaihtelusta. Esimerkkitapauksessamme jäljellä ovat enää Auran ja Vihdin sanaston eroavaisuudet, ja niinpä kuvassa 2 toisella pääkomponentilla vain Aurassa esiintyvä *kräki* erottuu kahdesta muusta länsisuomalaisesta sanasta. Esimerkkitapauksessamme koko aineisto on kuvattavissa vain näiden kahden komponentin avulla. Yleisemmässä tapauksessa voisimme jatkaa tästä: kolmas pääkomponentti selittäisi sitä osaa vaihtelusta, joka kahdelta ensimmäiseltä pääkomponentilta jäi selittämättä, ja niin edelleen. Yhdessä kolme ensimmäistä pääkomponenttia antavat siis eräässä mielessä parhaan kuvauksen aineistosta, joka voidaan antaa vain kolmen muuttujan avulla.



Kuva 2. Sanojen sijoittuminen 1. ja 2. pääkomponentin suunnassa.

Pääkomponentit voidaan havainnollistaa kartalla, jossa näkyy, millä painolla kukin pitäjä on mukana kyseisessä komponentissa. Komponentin ääripäät on piirretty mustana ja valkoisena, ja harmaan eri sävyillä piirretyillä kunnilla on vähemmän painoa komponentissa. Yllä kuvatussa esimerkissämme tämä tarkoittaisi sitä, että ensimmäisen pääkomponentin kartalla Aura olisi valkoinen, Kitee ja Juva olisivat mustia ja Vihti olisi vaalenharmaa. Toisen pääkomponentin kartalla Vihti olisi valkoinen, Aura musta ja Kitee ja Juva harmaita. Se, kumpi ääripää on valkoinen ja kumpi musta, ei ole merkityksellistä. Mielekästä on tarkastella, mihin ääripäät sijoittuvat: tämä kertoo kyseisen pääkomponentin selittämästä vaihtelusta. Juuri tällaisesta havainnollistamisesta on kyse kuvissa 5–9 (s. 42–44), joissa pääkomponentit on laskettu koko aineistostamme.

SANASTOPOHJAINEN MURREJAKO

Sanastoon perustuva, ryvästämällä tehty aluejako noudattelee huomattavan hyvin, muttei kuitenkaan täysin, perinteistä murrejakoa. Rapolan tuntuma osoittautui siis oikeaksi: sanaston perusteella voidaan saada lisävalaistusta suomen murteistoon, vaikka päälinjat onkin jo selvitetty äänne- ja muotopiirteiden perusteella. Tätä käsitystä vahvistaa se, että sanastoon perustuva raja poikkeaa perinteisestä lähinnä sellaisissa kohdissa, joissa Wiikin (2004) laskemat murrerajat ovat tietyn alueen ympärillä osapuilleen yhtä voimakkaat kahteen eri suuntaan.

Näissä muutamassa tapauksessa äänne- ja muotopiirteisiin perustuva murreraja voitaisiin vetää lähes yhtä luontevasti kahdesta eri kohtaa, ja valinta on vanhastaan tehty kieli- ja asutushistoriallisen tiedon perusteella. Sanaston pohjalta sen sijaan voi todeta, että »nykyhetken» näkökulmasta — jos yli puolen vuosisadan takaista tilannetta voi sanoa nykyhetkeksi — perinteinen valinta on väärä, ja murreraja oikeastaan kulkee luontevammin vaihtoehtoista linjaa pitkin.

▷

Pohjoinen murrealue

Suurin ero vakiintuneeseen murrejakoon on, että kahtiajako itä- ja länsimurteisiin sopii aineistoon kovin huonosti. Toki jaettaessa aineistoa kahtia raja kulkee melko hyvin perinteistä rajalinjaa pitkin, kuten kuvasta 4 (s. 42) näkyy, vaikkakin Keuruun–Evijärven alue sijoittuu länsi- ja valtaosa Kymenlaaksoa itämurteiden puolelle. Kun alueet jaetaan kolmeen osaan, tilanne kuitenkin muuttuu radikaalisti, ja jakoa myöhemmin tihennettäessä nimenomaan kolmijaon rajalinjat osoittautuvat pysyviksi.

Kahden murreryhmän sijasta on siis selvästi luontevampaa jakaa suomen murteet kolmeen osaan, itä-, länsi- ja pohjoismurteisiin. Ajatus ei sinänsä ole uusi: lähes nyt esittämämme kaltaista kolmijakoa on viimeksi tarjonnut Paunonen (1991) ja jo paljon varhemmin Warelius (1848). Onkin hiukan surullista, että Warelius on vanhastaan nähty uranuurtajana, joka kartoitti itä- ja länsimurteiden välistä rajaa, ja samalla unohdettu, että hän hiukan jäljempänä samassa artikkelissaan nosti näiden rinnalle vielä Pohjanmaalla puhutun »Kainuun murteen».[1]

Wareliuksen Kainuun murretta ei pidä sekoittaa nykyiseen samannimiseen murrealueeseen: Warelius itse laski Kajaanin murteen — siis sen, jota nykyisin kutsutaan Kainuun murteeksi — kuuluvaksi Itä-Suomen eikä Kainuun — siis Pohjanmaan — murteisiin. Tässä kohden poikkeamme jälleen verraten paljon perinteisestä murrejaosta: sanastoon perustuva jakomme nimittäin liittää Kainuun murteet selvääkin selvemmin osaksi pohjoisia murteita. Tämä on ehkä näkyvin kohta, jossa perinteisen murrejaon linjaus on tehty asutushistorian perusteella: äänne- ja muotopiirteiden perusteella vedetty raja on Pohjois-Pohjanmaan suuntaan aivan yhtä häilyvä ja Peräpohjolaankin päin vain hiukan voimakkaampi kuin Savon suuntaan (Wiik 2004: 25–27).

Sanastoltaan Kainuun murteet sen sijaan kuuluvat selvästi pohjoiseen murreryhmään ja siellä osaksi Pohjanmaan murteita. Raja Pohjois-Pohjanmaan ja Kainuun välillä on heikko ja tulee näkyviin vasta verraten myöhään; lopulta ilmaantuessaan se kulkee Ranuan ja Pudasjärven länsirajaa, kuten Räisänen (1972: 21) on esittänyt.[2] Ryvästyksemme vetää myös Pohjois-Pohjanmaan murteiden etelärajan selvästi etelämmäksi kuin perinteisesti, vanhalle Oulun läänin etelärajalle. Myös Wiik esittää tähän kohtaan voimakkaamman murrerajan kuin perinteiselle Raahen korkeudella kulkevalle linjaukselle, ja jo Rapola (1961: 103) piirtää tähänkin kohtaan murrealueen sisäisen rajan.

Hämäläis- ja lounaismurteiden välinen alue

Pohjanmaan lisäksi eroja perinteiseen murrejakoon löytyy myös Hämeen ympäristöstä. Kuten tunnettua, hämäläismurteiden reunoilla on niin savolais- kuin lounaismurteidenkin suunnalla siirtymämurteita. Tämä näkyy selvästi myös sanastossa, mutta tutkimuksemme sijoittelee monet näistä murteista eri tavalla kuin vanhastaan on totuttu tekemään.

---

[1] Todettakoon tässä myös, että kahtiajaon ensimmäiseksi esiintymäksi mainittu Vhaelin (1733) kielioppi ei sekään täysin pitäydy tässä jaossa: »päämurteen» ja Savon murteen lisäksi siinä mainitaan tarpeen tullen myös Pohjanmaan ja Karjalan murteet.

[2] Räisänen vetää murrerajan Ranuan keskeltä, mutta omassa aineistossamme aluejako on tarkkuudeltaan karkeampi, vain pitäjittäinen.

Hämäläis- ja lounaismurteiden välistä on noin neljän vuosikymmenen ajan totuttu erottamaan lounaiset välimurteet omaksi alueekseen. Kyseessä on kylläkin nuorin murrealueistamme, eivätkä Kettunen (1940) ja Rapola (1961) sitä vielä tunne, mutta sen sijaan Itkonen (1965: 31) esittää sen jo muiden rinnalla. Alueen nostaminen tällä tavoin esille on sinänsä perusteltua, jos tilannetta tarkastellaan äänne- ja muoto-opillisten piirteiden kannalta: murreraja lounaismurteiden suuntaan on erityisesti alueen pohjoisosassa hyvinkin jyrkkä, ja myös raja hämäläismurteisiin on selvästi nähtävissä (Wiik 2004: 24–27). Sanastoa tarkasteltaessa tilanne ei kuitenkaan ole yhtä selvä.

Ryvästys ei missään vaiheessa erota lounaisten välimurteiden aluetta omaksi rypäkseen, vaan sen sijaan alue liittyy läheisesti lounaismurteisiin. Kun sanaston pohjalta laadittua jakoa riittävästi tihennetään, Porin seudun ja Ala-Satakunnan alueet yhdistyvät lopulta hämäläismurteiseksi vanhastaan laskettuun Ylä-Satakuntaan ja muodostavat selvärajaisen satakuntalaisen alueen. Tällaisen alueen olemassaoloa puoltaa sanastollisesti myös kuvassa 9 (ks. s. 44) näkyvä 8. pääkomponentti, jonka ydinalue on selvän satakuntalainen. Samaten on huomattava, että vaikka Ylä-Satakunnan länsireunalla on selvä äänne- ja muotopiirteisiin perustuva murreraja, se on merkittävästi heikompi kuin satakuntalaisen alueen ja lounaismurteiden välinen.

Länsi-Uudenmaan ja Turun ylämaan alueet puolestaan jäävät edelleenkin lounaismurteiden yhteyteen. Kaiken kaikkiaan sanasto viittaisi siihen, että Rapola (1961) oli oikeilla jäljillä esitellessään nämä murteet lounaismurteiden, mutta Porin ja Ala-Satakunnan murteet hämäläismurteiden yhteydessä. Myös Suomen murteiden sanakirja (Tuomi 1989: 94) on samoilla linjoilla, vaikkakin tuossa tapauksessa ryhmittely perustuu murteiden lisäksi paljolti maakuntien nimiin. Lounaisten välimurteiden olemassaolon puolesta on toisaalta huomattava niiden voimakas äänne- ja muoto-opillinen murreraja lounaismurteiden suuntaan.

Itämurteiden länsiraja

Jos hämäläismurteiden länsiraja on epäselvä, myös raja itään on häilyvä. Vanhastaan savolaismurteisiin luetuissa Päijät-Hämeen ja Keski-Suomen murteissa on tunnetusti paljon hämäläistäkin, ja tämä näkyy myös sanastossa. Päijät-Häme on näistä kahdesta hiukan hämäläisempi ja Keski-Suomi savolaisempi alue, mikä ei myöskään ole mitenkään uusi havainto.

Savolaismurteiden länsireunalla poikkeamme perinteisestä luokittelusta selvimmin Keuruun–Evijärven välimurteiden kohdalla. Erityisesti alueen pohjoisosan ympärillä on joka puolella hyvin voimakas murreraja (Wiik 2004: 41), ja savolaismurteisiin se lienee luokiteltu lähinnä asutushistoriansa perusteella; niinpä jo Rapola (1961: 135) luonnehtii aluetta vain puoliksi savolaiseksi. Sanaston tarkastelu kallistaa kuitenkin vaakaa toiseen suuntaan, ja alueen pohjoisosa näyttäisi olevan pikemminkin osa Keski-Pohjanmaan murteistoa; eteläisin osa sen sijaan asettuu osaksi Päijät-Hämeen ja Keski-Suomen länsisavolaista murrealuetta.

Wiik (2004: 45) pitää aluetta savolaismurteisiin kuuluvana, lähinnä koska alueen raja Etelä-Pohjanmaan suuntaan on voimakkain suomen murteista löytyvä, eikä sitä siksi ole mielekästä tulkita kahden länsimurteen väliseksi. Omassa jaottelussamme tätä ongelmaa ei kuitenkaan ole, koska kyseessä on kahden pääryhmän, länsi- ja pohjoismurteiden, väli- ▷

nen raja. Toki on myönnettävä, että alue on kaiken kaikkiaan selvin esimerkki siitä, kuinka vaikeaa siirtymämurteiden luokittelu jo lähtökohtaisesti on.

Etelämpänä huomio kiinnittyy kaakkoishämäläisiin murteisiin. Niiden ympärillä kulkee verraten voimakas raja niin hämäläis- kuin kaakkoismurteidenkin suuntaan (Wiik 2004: 41), ja myös alueen sisällä Hollolan–Porvoon ja Kymenlaakson ryhmän välillä raja on jyrkkä. Tutkimuksemme esittää kuitenkin alueen yhtenäisenä ja lähinnä hämäläismurteisiin kuuluvana, vaikka mahdotonta ei olisi myöskään ajatella, että ne ryhmiteltäisiin Keski-Suomen ja Päijät-Hämeen kanssa yhteen jonkinlaisiksi hämäläis–itäisiksi siirtymämurteiksi.

Pääpiirteittäinen murrejako

Aineistomme perusteella ei voi kaikissa tapauksissa varmasti sanoa, mihin alueeseen tietty pitäjä kuuluu. Tämä johtuu osittain käytetyistä ryvästysmenetelmistä, mutta ennen kaikkea eri pitäjien kovin erilaisesta sanastusasteesta. Tulokset ovat tosin vielä melko luotettavia, vaikka yksittäisestä pitäjästä olisikin vähän kerättyä sanastoa, mutta jos koko murrealue on jäänyt vähälle huomiolle — kuten on laita ennen kaikkea Suomen talvisotaa edeltävien rajojen ulkopuolella — epätarkkuus kasvaa olennaisesti (Hyvönen ym. tulossa). Niinpä murteiden väliset rajat on edellä kuvattu pääasiassa perinteisen murrejaon rajojen avulla, vaikka kuvan 3 kartassa rajat menevät joiltakin osin hiukan eri kohdissa. Tällaisin varauksin esitämme, että suomi jakaantuu sanastonsa perusteella synkronisesti tarkasteltuna luontevimmin seuraavasti:

Länsimurteet

1. Lounaismurteet
   a) Pohjoisryhmä, kuten perinteisessä murrejaossa
   b) Itäryhmä sisältää myös Turun ylämaan, Someron ja Länsi-Uudenmaan murteet
2. Hämäläismurteet
   a) Satakunnan murteet sisältävät Ylä-Satakunnan lisäksi myös Porin seudun ja Ala-Satakunnan murteet
   b) Keskiset hämäläismurteet sisältävät peri- ja etelähämäläiset murteet
   c) Kaakkoishämäläiset murteet
3. Etelä-Pohjanmaan murre

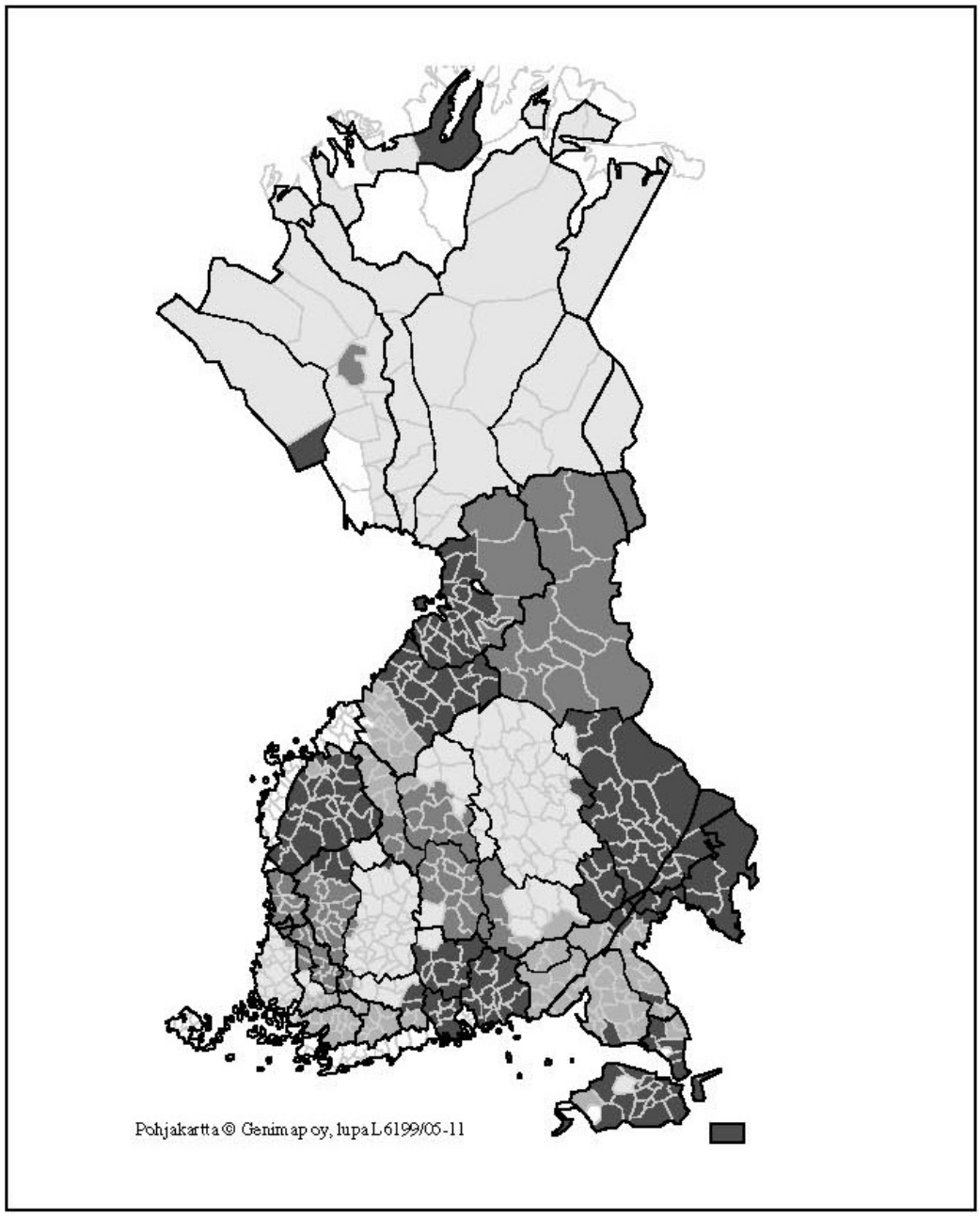
Itämurteet

4. Savolaismurteet
   a) Länsisavolaiset murteet sisältävät Keski-Suomen ja Päijät-Hämeen murteet sekä osan Keuruun seudun välimurteista
      (Kartan kaakkoisnurkkaan piirretyn suorakaiteen osoittama Vermlannin murre asettuu useimmissa ryvästyksissä tähän ryhmään, joskus myös Inkerin murteiden yhteyteen.)
   b) Sydänsavolaiset murteet sisältävät pohjois- ja eteläsavolaiset murteet
   c) Itäsavolaiset murteet sisältävät Pohjois-Karjalan murteet sekä Savonlinnan seudun välimurteet
5. Kaakkoismurteet
   a) Suomen kaakkoismurteet
   b) Inkerin suomalaismurteet
      Näiden välinen raja kulkee sotia edeltänyttä valtakunnanrajaa pitkin.

Pohjoismurteet

6. Pohjanmaan murteet
   a) Keski-Pohjanmaan murteet sisältävät vanhan Vaasan läänin pohjoisosan Ähtärin–Evijärven seudun välimurteista alkaen
   b) Pohjois-Pohjanmaan murteet sisältävät Oulun lääniin kuuluneen osan Pohjanmaan rannikkoa
   c) Kainuun murteet viimeaikaisen tulkinnan mukaan, Ranua ja Pudasjärvi mukaan luettuina
7. Peräpohjolan murteet

Kuva 3. Jako viiteentoista alueeseen.



Vaikka ryvästämällä onkin mahdollista muodostaa murrealueita, tämä ei kuitenkaan lopulta ole paras lähestymistapa, vaan hedelmällisempää on tarkastella pääkomponentteja suoraan. Murrevaihtelua koskevat paljolti samanlaiset havainnollistus- ja esitystapakysymykset, joista Haspelmath (2004: 216) mainitsee Australian ja Afrikan kieliolojen yhteydessä: hierarkkisen suku- tai muun senkaltaisen puun sijasta on luontevampaa esittää eri ilmiöiden alueellista diffuusiota karttoina. Kun lisäksi murrealueiden määrittämisessä on kyse siitä, että murteiden jatkumo jaetaan erillisiin osiin — vieläpä jossain määrin mielivaltaisesti, kuten Rapolakin totesi —, pääkomponenttien avulla tarkastelu voidaan kohdistaa tähän jatkumoon. Itse asiassa tilanne on vielä tätäkin parempi, koska kuhunkin komponenttiin on eristetty yksi tämän vaihtelun suunnista.

Hiukan samantapaista lähestymistapaa ovat perinteisemmän dialektometrian keinoin kokeilleet myös Heeringa ja Nerbonne (2002). Heidän menetelmällisenä lähtökohtanaan on kuitenkin murreasujen Levenštein-etäisyys, joka tuo mukanaan tiettyjä rajoituksia. Vaikka tutkimus pystyy jakamaan hollannin murteet muutamaan toisiaan verraten lähellä olevaan ryhmään, siinä kahden murteen välinen etäisyys on aina sama. Pääkomponenttianalyysin avulla suomen murteista löytyy useita vaihtelusuuntia, ja kaksi murretta voivat olla yhden komponentin suhteen lähellä toisiaan mutta toisen komponentin valossa hyvinkin etäisiä. Myös Wiikin (2004) murrepitoisuuskarttojen voidaan ajatella esittävän murrevaihteluiden jatkumoa. Kyseessä on kuitenkin sikäli erilainen lähestymistapa, että Wiik ottaa lähtökohdakseen olemassa olevan murrejaon ja tarkastelee kunkin alueen leimallisten piirteiden esiintymistä. Pääkomponenttianalyysi sen sijaan etsii aineistosta voimakkaimmat vaihtelusuunnat ilman tällaisia ennakkoaavistuksia.

Sanastosta löytyvistä komponenteista ensimmäinen (kuva 5, s. 42) osoittaa jokseenkin suoraan, kuinka kattavasti pitäjä on sanastettu, ja myös seitsemäs komponentti (kuva 8, s. 44) korreloi varsin voimakkaasti sanastusasteen kanssa. Tämä ei sinänsä ole lainkaan kiinnostavaa, etenkään kun kerättyjen sanojen pitäjittäiset lukumäärät tunnetaan muutenkin, ja edellä esitettyä murrejakoa laatiessamme olemme varsinaista ryvästystä edeltäneessä esikäsittelyvaiheessa pyrkineet poistamaan sen vaikutuksen. Sanastusaste on toisaalta kuitenkin sikäli hyödyllinen tieto, että se antaa osviittaa muiden tulosten luotettavuudesta: jos pitäjästä on saatavilla vain vähän tietoja, niiden perusteella tehdyt päätelmät eivät ole yhtä luotettavia kuin kattavammin kerätyn sanaston perusteella tehdyt.

Kuvassa 6 (s. 43) näkyvä toinen komponentti — siis ensimmäinen varsinaisesta murrevaihtelusta eristetty — osoittaa selvästi itä–länsi-vaihtelua. Läntinen ääripää löytyy odotuksenmukaisesti lounaismurteiden alueelta, itäinen puolestaan savolaismurteiden itäiseltä äärilaidalta Pohjois-Karjalasta. Komponentti ei kuitenkaan ole niin selvärajainen, että sillä voisi perustella jo Rapolan (1961: 106) esittämää ajatusta suomen murteiden pääjaon piirtämiseksi lounaismurteiden ja muiden väliin.

Kolmas komponentti puolestaan kulkee pohjois–eteläsuunnassa. Merkille pantavaa on, että lähes koko pohjoinen alue asettuu itä–länsi-vaihtelussa vaihteluvälin keskelle, mikä tukee ajatusta kolmesta päämurreryhmästä. Poikkeuksena tästä on Kainuu, etenkin sen eteläisimmät pitäjät; toisaalta juuri eteläisessä Kainuussa myös pohjoinen komponentti on melko voimakas. Eteläinen ääripää löytyy kaakkoismurteiden alueelta, pohjoinen puolestaan on hajanaisempi ilmeisesti lähinnä pohjoisimpien kuntien pienen sanastusasteen vuoksi.

Näiden alaltaan maan- tai kielenlaajuisten komponenttien jälkeen on vuorossa joukko maantieteellisesti rajatumpia ilmiöitä. Niistä ensimmäisenä esiin nousee neljäs komponentti, kuvassa 7 (s. 43), joka lienee tulkittava lähinnä hämäläiseksi: keskeinen alue kulkee Kymenlaaksosta kaakkois- ja pohjoishämäläisen alueen läpi Etelä-Pohjanmaalle saakka. Tämän laajan hämäläiskomponentin pariksi on hyvä huomata vielä kuvassa 9 (s. 44) näkyvä kahdeksas komponentti, jonka ala rajoittuu selvemmin Satakuntaan.

Hämeen jälkeen seuraavana nousee esille Etelä-Pohjanmaa. Kuvassa 7 näkyvä viides komponentti on lisäksi jyrkkärajaisempi kuin muut aineistostamme löytyneet, mikä tukee ajatusta eteläpohjalaisen murrealueen selvästä erityisluonteesta. Tämä on erityisen ilmeistä, jos sitä verrataan kuvan 8 lähinnä savolaisuuteen liittyvään kuudenteen komponenttiin, joka jatkuu ydinalueiltaan Pohjois-Savosta ja Pohjois-Karjalasta laimentuneena aina länsimurteisiin saakka.

LOPUKSI

Kaiken kaikkiaan sanastoanalyysimme jakaa suomen murteet tavalla, joka enimmäkseen vastaa jo vakiintunutta käsitystä. Kun lähtökohta on näinkin kaukana perinteisestä, tätä tulosta on pidettävä varsin voimakkaana riippumattomana osoituksena siitä, että vanha murrejako on hyvin perusteltu ja oikeaan osunut. Kuitenkin myös eroja löytyy, ja nämä ovat pääosin sellaisissa kohdin, joissa äänne- ja muotopiirteiden perusteella raja olisi voitu vetää lähes yhtä hyvin kahteen eri kohtaan. Oma jakomme on paremmin perusteltu, kun kieltä tarkastellaan synkronisesti; perinteistä jakoa puolestaan voi pitää diakronisesti sopivampana. On myös syytä huomata, että murrejako antaa lopulta melko yksipuolisen kuvan murteista: kahden murteen välinen ero on moniulotteisempi ilmiö kuin mitä tällaisella kartalla voi esittää. Aluejaon lisäksi onkin syytä kiinnittää huomiota siihen, että murteista löytyy erilaisia vaihtelusuuntia.

Tällaisista vaihteluista merkittävimmät on esitetty kuvien 6–9 kartoissa (s. 43–44). Ne heijastuvat toki myös murrejaossa: kun voimakkain murteiden vaihtelusta eristetty komponentti noudattelee perinteistä itä–länsi-jakoa ja seuraava pohjois–etelä-jakoa, ne yhdessä selittävät jaon itä-, länsi- ja pohjoismurteisiin. Näiden lisäksi on muitakin vaihtelusuuntia, jotka nostavat esiin alueellista, esimerkiksi hämäläistä tai savolaista vaikutusta. Jotkin murrevaihtelun komponenteista voivat jopa erottaa tietyn alueen hyvin selvärajaisesti muusta murteistosta; näin käy jo varhaisessa vaiheessa Etelä-Pohjanmaalle.

Jatkossa olisi kiinnostavaa päästä tekemään tässä esitetyn kaltainen analyysi myös niille murrepiirteille, joiden perusteella perinteinen murrejako on laadittu. Tämänsuuntaisia ajatuksia on esittänyt ainakin Palander (1999); Wiik (2004) puolestaan kertoo esipuheessaan, että aineistoa olisi jo jossain määrin tietokonemuodossa, ja myös Embleton ja Wheeler (2000) mainitsevat aineistonsa olevan pääosin valmiina. Joka tapauksessa jo tässä vaiheessa näyttää selvältä, että tietotekniikan kehittyminen avaa uusia näkökulmia myös perinteisiin kielentutkimuksen aloihin. Samoin on ilmeistä, että sanaston kautta on saatavissa lisävalaistusta murteiden vaihteluun.

LÄHTEET

BERRY, MICHAEL W. – DRMAC, ZLATKO – JESSUP, ELISABETH R. 1999: Matrices, vector spaces, and information retrieval. – *SIAM Review* 41(2) s. 335–362.

EMBLETON, SHEILA – WHEELER, ERIC S. 1997: Finnish dialect atlas for quantitative studies. – *Journal of Quantitative Linguistics* 4(1–3) s. 99–102.

——— 2000: Computerized dialect atlas of finnish: Dealing with ambiguity. – *Journal of Quantitative Linguistics* 7(3) s. 227–231.

GOOSKENS, CHARLOTTE– HEERINGA, WILBERT 2004: Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. – *Language Variation and Change* 16 s. 189–207.

HAND, DAVID – MANNILA, HEIKKI – SMYTH, PADHRAIC 2001: *Principles of data mining*. Cambridge: The MIT Press.

HASPELMATH, MARTIN 2004: How hopeless is genealogical linguistics, and how advanced is areal linguistics? – *Studies in Language* 28(1) s. 209–223.

HEERINGA, WILBERT – NERBONNE, JOHN 2002: Dialect areas and dialect continua. – *Language Variation and Change* 13 s. 375–398.

HOTELLING, HAROLD 1933: Analysis of a complex of statistical variables into principal components. – *Journal of Educational Psychology* 24 s. 417–441, 498–520.

HYVÖNEN, SAARA – LEINO, ANTTI – SALMENKIVI, MARKO tulossa: *Multivariate analysis of Finnish dialect data*.

ITKONEN, TERHO 1965: *Proto-Finnic final consonants*. Suomalais-Ugrilaisen Seuran toimituksia 138:1. Helsinki: Suomalaisen Kirjallisuuden Seura.

KETTUNEN, LAURI 1940: *Suomen murteet III A. Murrekartasto*. Helsinki: Suomalaisen Kirjallisuuden Seura.

LEVENSHTEIN, VLADIMIR IOSIFOVITCH 1966: Binary codes capable of correcting deletions, insertions, and reversals. – *Soviet Physics Doklady* 10 (8) s. 707–710.

MACQUEEN, JAMES 1967: Some methods for classification and analysis of multivariate observations. – Lucien M. Le Cam & Jerzy Neyman (toim.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* I s. 281–297. Berkeley: University of California Press.

MIELIKÄINEN, AILA 1990: Savolais- ja kaakkoismurteiden rajankäyntiä. *Laatokan piiri. Juhlakirja Heikki Leskisen 60-vuotispäiväksi 10.10.1990* s. 112–130. Kotimaisten kielten tutkimuskeskuksen julkaisuja 60. Helsinki: Kotimaisten kielten tutkimuskeskus.

——— 1991: *Murteiden murros. Levikkikarttoja nykypuhekielen piirteistä*. Jyväskylän yliopiston suomen kielen laitoksen julkaisuja 36. Jyväskylä: Jyväskylän yliopisto.

NERBONNE, JOHN 2003: Linguistic variation and computation. – *Proceedings of the 10th Meeting of the European Chapter of the Association for Computational Linguistics* s. 3–10.

NERBONNE, JOHN – HEERINGA, WILBERT 2001: Computational comparison and classification of dialects. – *Dialectologia et Geolinguistica* 9 s. 69–83.

PALANDER, MARJATTA 1996: *Vaihtelu Savonlinnan seudun välimurteissa*. Suomalaisen Kirjallisuuden Seuran toimituksia 648. Helsinki: Suomalaisen Kirjallisuuden Seura.

——— 1999: Mitä dialektometria on? – *Virittäjä* 103 s. 259–265.

Palander, Marjatta – Opas-Hänninen, Lisa Lena – Tweedie, Fiona 2003: Neighbours or enemies? Competing variants causing differences in transitional dialects. – *Computers and the Humanities* 37 s. 359–372.

Paunonen, Heikki 1991: Till en ny indelning av de finska dialekterna. – *Fenno-Ugrica Suecana* 10 s. 75–79.

Pearson, Karl 1901: On lines and planes of closest fit to systems of points in space. – *London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 6 (2) s. 559–572.

Rapola, Martti 1961: *Johdatus suomen murteisiin*. Toinen, uudistettu painos. Tietolipas 4. Helsinki: Suomalaisen Kirjallisuuden Seura.

Ruoppila, Veikko 1937: Kaakkois-suomen murteista. – *Virittäjä* s. 58–71.

Räisänen, Alpo 1972: *Kainuun murteiden äännehistoria. I: Vokaalisto*. Suomalaisen Kirjallisuuden Seuran toimituksia 307. Helsinki: Suomalaisen Kirjallisuuden Seura.

Savijärvi, Ilkka 1995: Hämäläismurteen murtuminen Päijänteen länsirannalla. – Marjatta Palander (toim.), *Murteiden matkassa. Juhlakirja Alpo Räisäsen 60-vuotispäiväksi*. Studia Carelica Humanistica 6. Joensuun yliopiston humanistinen tiedekunta.

Savijärvi, Ilkka – Yli-Luukko, Eeva 1994: *Jämsän äijän murrekirja*. Suomalaisen Kirjallisuuden Seuran toimituksia 618. Helsinki: Suomalaisen Kirjallisuuden Seura.

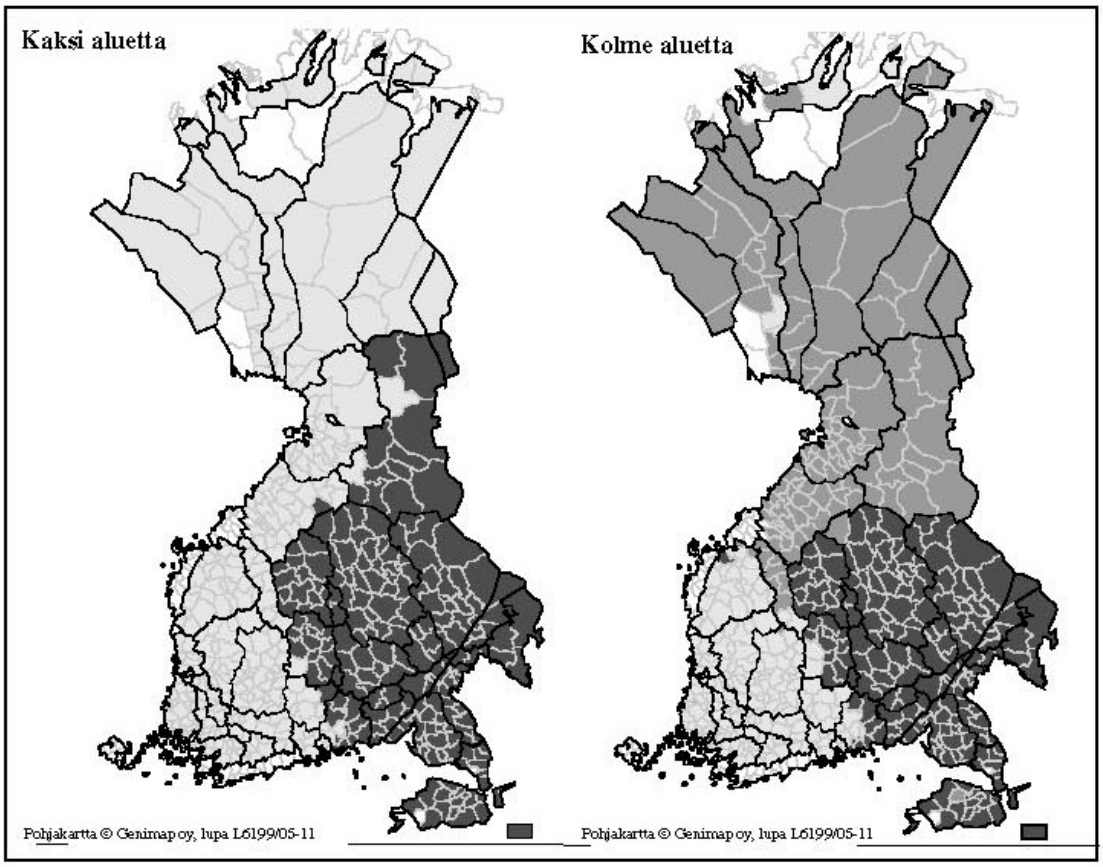
Strandberg, Jan 2004: *Ei sanat salahan joua: fennistiikan murteenkeruun historiaa 1868–1925*. Pro gradu -tutkielma. Helsingin yliopiston suomen kielen laitos.

Tuomi, Tuomo (toim.) 1989: *Suomen murteiden sanakirja. Johdanto*. Kotimaisten kielten tutkimuskeskuksen julkaisuja 36. Helsinki: Kotimaisten kielten tutkimuskeskus.

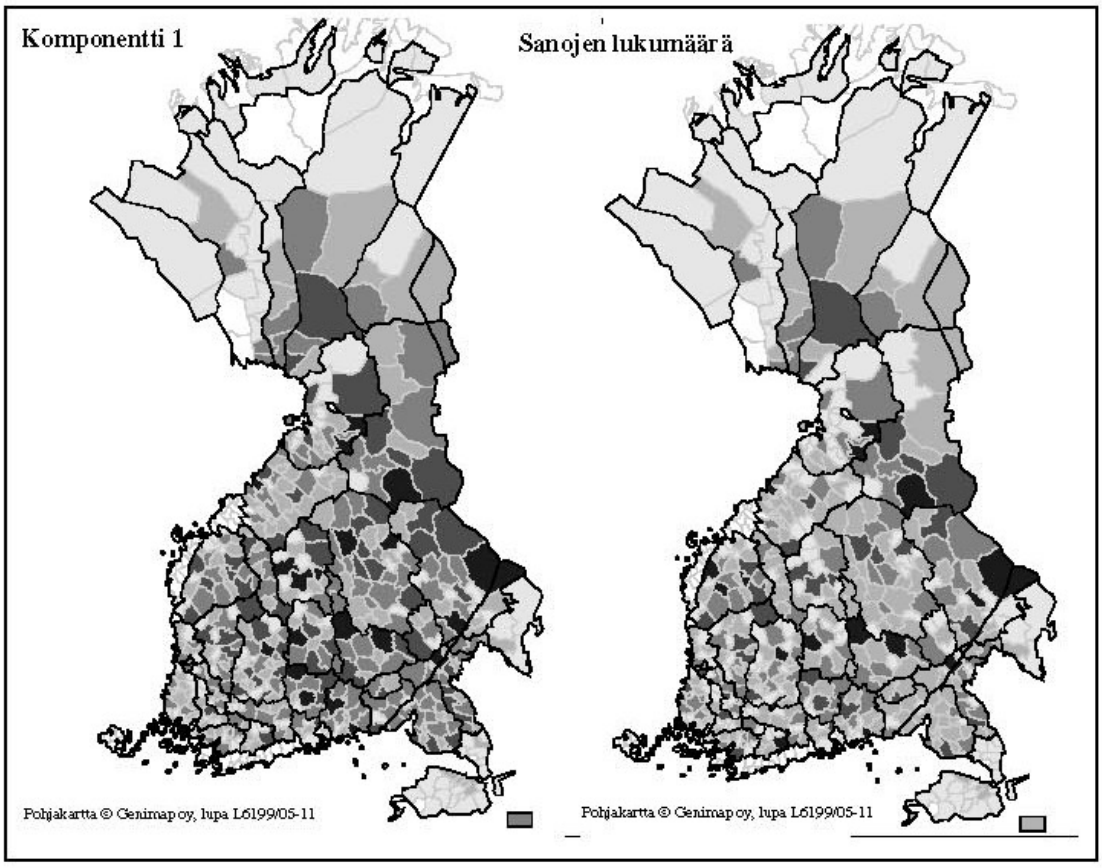
Vhael, Bartholdus G. 1733: *Grammatica Fennica*. Johan Kiämpe. Näköispainos: *Vanhat kielioppimme.* Suomalaisen Kirjallisuuden Seura 1968.

Warelius, Anders 1848: Bidrag till Finlands kännedom i ethnographiskt hänseende. – *Suomi* 7 s. 47–130.

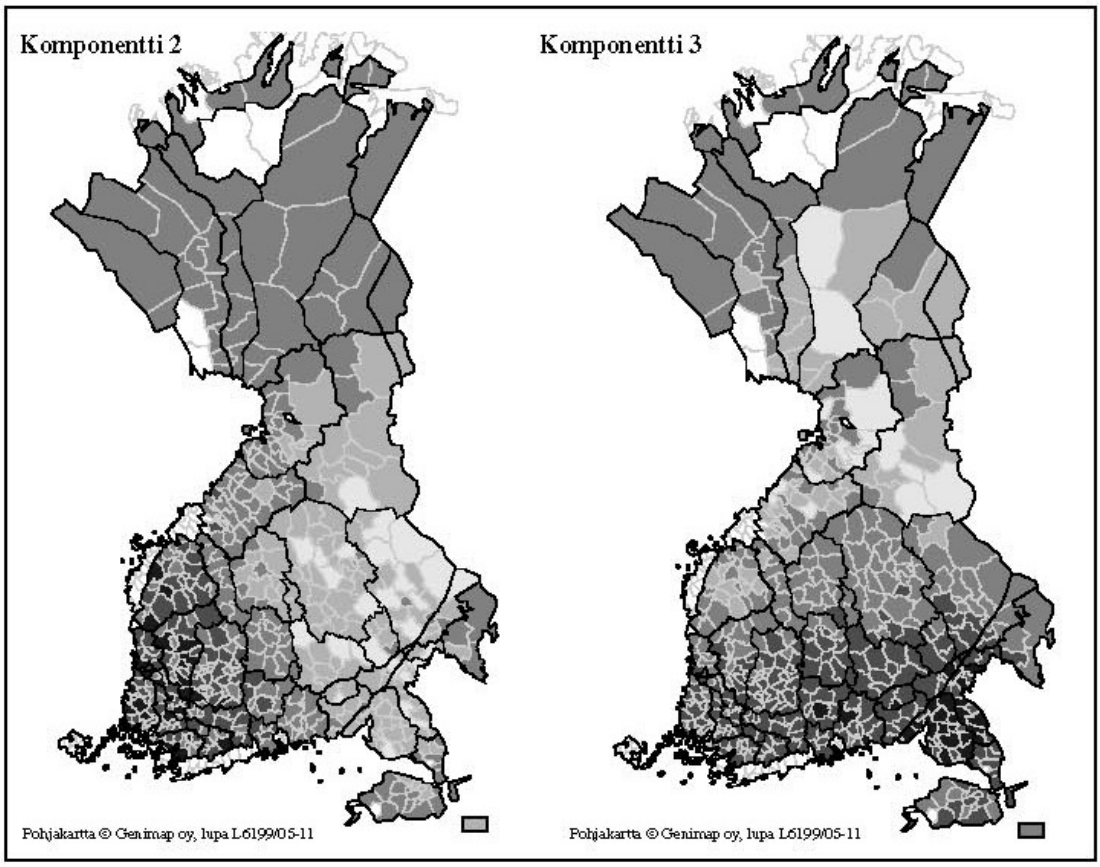
Wiik, Kalevi 2004: *Suomen murteet. Kvantitatiivinen tutkimus*. Suomalaisen Kirjallisuuden Seuran toimituksia 987. Helsinki: Suomalaisen Kirjallisuuden Seura.



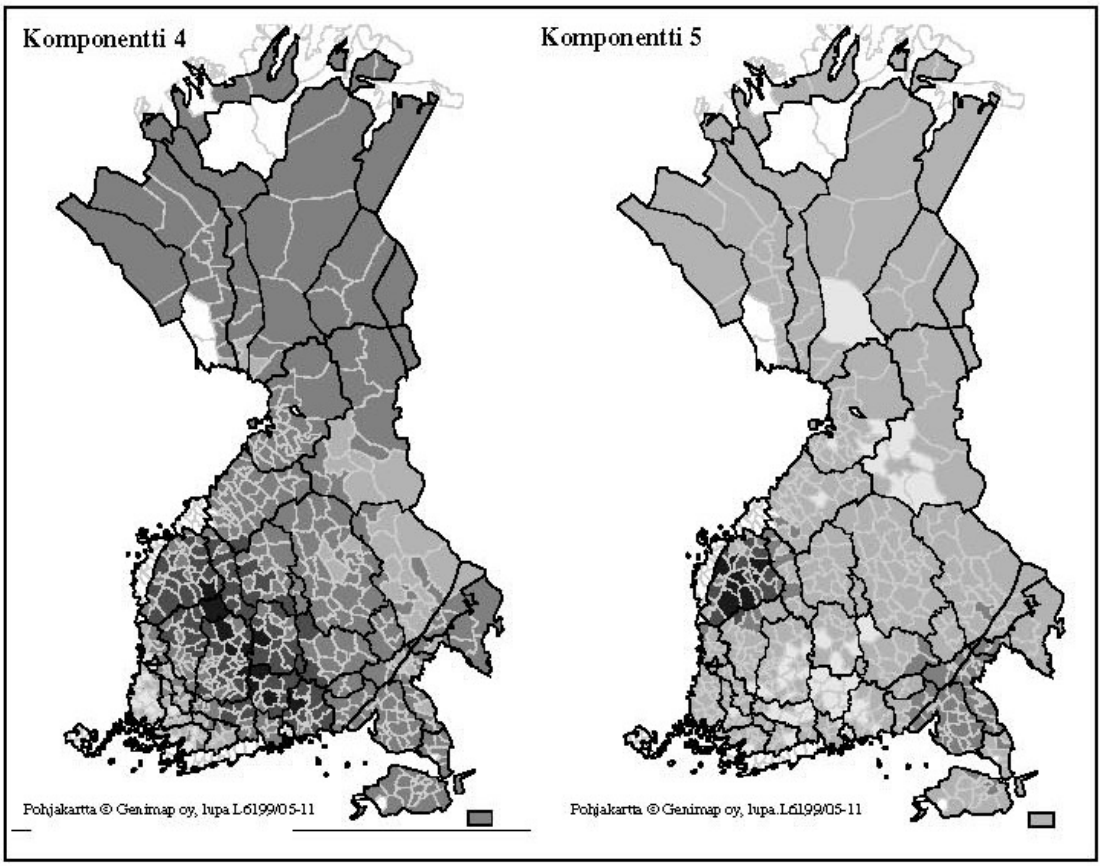
Kuva 4. Jako kahteen ja kolmeen alueeseen.



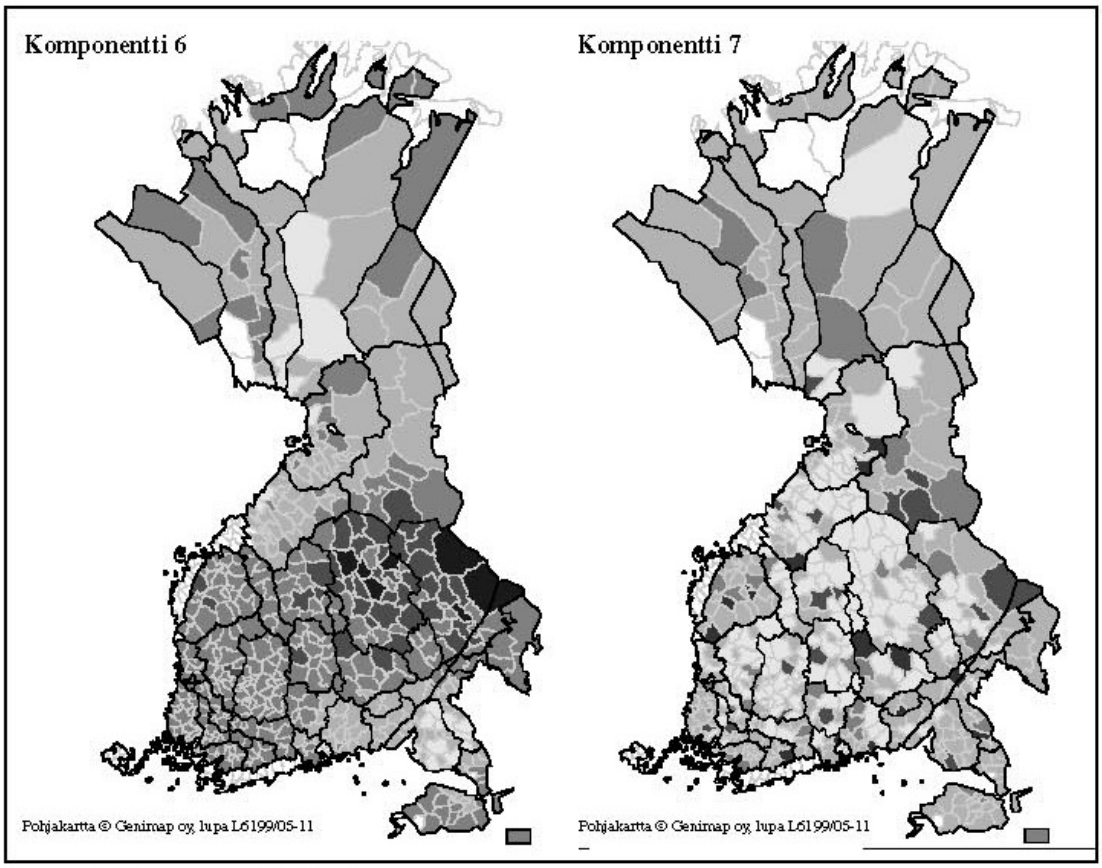
Kuva 5. Komponentti 1.



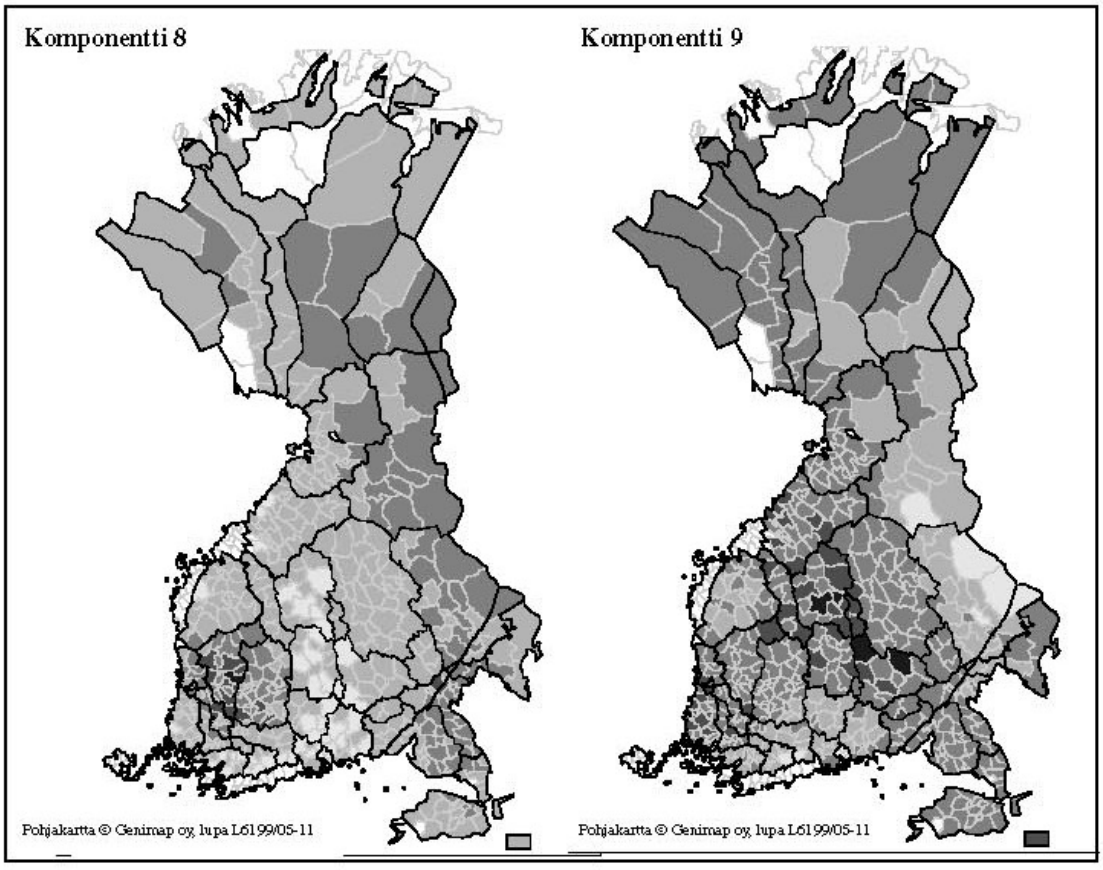
Kuva 6. Komponentit 2 ja 3.



Kuva 7. Komponentit 4 ja 5.



Kuva 8. Komponentit 6 ja 7.



Kuva 9. Komponentit 8 ja 9.

A QUANTITATIVE ANALYSIS OF THE DISTRIBUTION OF DIALECT WORDS

Finnish dialects have traditionally been divided into eastern and western dialects, although a three-way division into eastern, western and northern dialects has sometimes also been proposed. In the established two-way division, the western dialects have been further divided into six main areas and the eastern dialects into two, and there has been considerable unanimity on the boundaries of these areas. This dialect division is based principally on phonological and morphological features.

The writers had access to some 9,000 different maps showing the distribution of dialect words. These maps were drawn up as part of the project to compile a dictionary of Finnish dialects. The writers have analysed this data using various data analysis tools, and the results so far have been fairly consistent with the previous understanding of the situation. The most significant difference, however, is that the analysis has strongly supported the idea that a northern dialects region can be just as clearly distinguished as the division between eastern and western dialects. This northern region contains the Central and Northern Ostrobothnian and Northern Finnish dialects, all previously classed as western dialects, and also the Kainuu dialects, which are traditionally considered a sub-group of the Savo dialects.

Areas and boundaries are not always the most informative way to illustrate the geographical variation in language, however. Instead, the extent of the differences between dialects can prove a rewarding area for study, and variation can be divided into separate, uncorrelated components. The writers present the most significant of these components, along with the traditional dialect maps.

The writers note that their analysis is primarily based on lexical considerations, which may affect the results a little. They also had access to only certain alphabetical sections of the dictionary. Nevertheless, the results support the traditionally held views quite closely, and the differences in comparison with previous studies are visible mainly in areas where phonological and morphological features provide scope for alternative interpretations. ■

Kirjoittajien yhteystiedot (addresses):
Tietojenkäsittelytieteen laitos
PL 68
00014 Helsingin yliopisto
Sähköpostit: *antti.leino@helsinki.fi*
saara.a.hyvonen@helsinki.fi
marko.salmenkivi@helsinki.fi