Kohti yhteisiä aineistokäytänteitä

Tutkimusaineistojen yhteiskäytöstä on digitaalisena aikana ja viimeistään 2000-luvulla tullut itsestään selvä tavoite kielentutkimuksen parissa, ja aineistoasiat ovat olleet yhä enemmän esillä niin Suomessa kuin kansainvälisestikin. *Virittäjässä* 3/2010 aineistoasioita käsiteltiin monipuolisesti Suomen kielen nauhoitearkiston 50-vuotisjuhlavuoden tapahtumien ja teemojen kautta. Lehden Suunvuoro-palstalla Toni Suutari (2010a) nosti esiin ajankohtaisen ja tärkeän kysymyksen kansallisen kieliaineistoinfrastruktuurin tarpeesta Suomessa ja kiinnitti samalla huomiota muun muassa tutkimusrahoittajien kiinnostukseen aineistojen jatkokäyttöä kohtaan. Erkki Lyytikäisen ja Jaakko Yli-Paavolan (2010) katsaus puolestaan valotti sitä, miten kielennäytteiden yhtenäisestä tallentamisesta on keskusteltu ja miten sitä on kehitetty nauhoitearkiston 50-vuotisen historian aikana. Tässä kirjoituksessa jatkan keskustelua kieliaineistojen arkistoinnista ja kiinnitän huomiota erityisesti kielentutkijoiden rooliin aineistokäytänteiden kehittäjinä.

Suomalaisen arkistoinfrastruktuurin on tulevaisuudessa toimittava yhä enemmän kansainvälisessä kontekstissa ja tuettava useiden eri kielten tutkimusta. Kansallisesti merkittävän aineistorakenteen profiili painottuu kuitenkin suomalais-ugrilaisiin ja Suomessa puhuttaviin kieliin ja niiden tutkimukseen. Kestävän aineistoinfrastruktuurin rakentaminen edellyttää riittävien resurssien ja teknisen toteutuksen lisäksi aineistonhallinnan käytänteiden suunnittelua ja lingvistiyhteisön, erityisesti fennistien ja fennougristien, vahvaa panosta. Uusien aineistojen tuottamisen, käsittelyn ja hallinnan yhteisiä käytänteitä ja suuntaviivoja tarvitaan, jotta yhteiskäyttöön soveltuvien aineistojen tuottaminen olisi tutkijoille vaivatonta ja luonteva osa tutkimusprosessia.

Aineistojen jakaminen säästää tutkimusyhteisön aikaa ja vaivaa, ja rinnakkain samanlaisten pienten aineistojen parissa tehtävän työn sijaan on mielekkäämpää keskittyä kartuttamaan eri aineistoista ja niiden analyyseistä muodostuvaa, edustavaa yhteiskorpusta. Suomessa etenkin pienempien suomalais-ugrilaisten kielten tutkimuksessa samojen klassikkoaineistojen hyödyntäminen aina uusissa tutkimuksissa on tuttua: esimerkiksi Suomalais-Ugrilaisen Seuran kentälle lähettämien stipendiaattien keräämät laadukkaat ja verrattain suuret aineistot ja niiden pohjalta 1900-luvun mittaan toimitetut kieliopit, tekstikokoelmat ja sanakirjat ovat olleet monilta osin suomalaisen fennougristiikan perusta (ks. lähemmin esim. Saarinen 2007; Grünthal 2010). Näitä klassikkoaineistoja käytetään edelleen pääaineistoina yksittäisten suomalais-ugrilaisten kielten tutkimuksessa, ja erityisesti kielten muu-

tosta heijastavina vertailuaineistoina niiden rooli on merkittävä myös tulevaisuudessa (Saarinen 2007). Vaikka 1800-luvun lopulta alkaen kerättyjä aineistoja on julkaistu läpi 1900-luvun ja digitoitu yhä enemmän, valitettavan pieni osa erityisesti aineistoihin eri tutkimusten yhteydessä tehdyistä kieliopillisista analyyseistä ja käännöksistä on muiden tutkijoiden ja kieliyhteisöjen saatavilla sähköisessä muodossa. Sukukielten lisäksi sama puute koskee monelta osin myös suomen kielen aineistoja.

Entistä kattavampien ja monipuolisemmin analysoitujen yhteisten aineistojen hyödyntäminen mahdollistaa lisäksi empiirisen tutkimuksen toistettavuuden. Tältä osin kielitiede lähestyy luonnontieteitä laadunvarmistuksen ja vertaisarvioinnin näkökulmasta. Vähemmän tutkittujen, pienten tai uhanalaisten kielten materiaalien pitkäjänteisesti suunniteltu arkistointi ja jakaminen voidaan nähdä myös työnä kielidiversiteetin tallentamiseksi ja esiin tuomiseksi, tutkimuksen lisäksi myös kielten revitalisaatio- ja opetushankkeiden käyttöön. Kieliteknologian tarjoamien sovellusmahdollisuuksien rajana ovat vain mielikuvitus ja määrärahat sekä aineistojen käyttöoikeudet.

Suomalainen arkistoinfrastruktuuri

Kieliaineistojen hallintaan ja tallentamiseen on eri maissa, tutkimuslaitoksissa ja yliopistoissa tarjolla monenlaisia työkaluja ja sähköisiä arkistoja. Sovellusten ja palveluiden runsaan määrän vuoksi haasteena onkin kieliaineistojen sekä eri arkistojen formaattien ja rakenteiden yhteensopivuus aineistojen siirrettävyyden takaamiseksi, jotta tietyllä sovelluksella tallennettu ja käsitelty aineistokokonaisuus olisi käytettävissä ja analysoitavissa myös muualla. Yhtä suuri haaste on taata tarjolla olevien palveluiden lähestyttävyys ja käytettävyys kieliaineistojen tallentajille ja tutkijoille, joista suurimmalla osalla ei ole kieliteknologista erityisosaamista.

Viime vuosina yhteiseurooppalainen Clarin-hanke (ks. http://www.clarin.eu) on rakentanut pohjaa kieliaineistojen yhteiskäytölle Euroopassa, ja sen kansallisena yhteistyötahona toimii Suomessa Fin-Clarin-konsortio. Yhteiseurooppalaisen hankkeen päämääränä on parantaa aineistojen ja niiden käyttöön ja käsittelyyn tarkoitettujen työkalujen saatavuutta yli instituutti- ja maarajojen esimerkiksi selkeyttämällä ja yhtenäistämällä aineistojen luettelointia ja käyttöluparatkaisuja. Suomalaisen konsortion ja siihen liittyvien hankkeiden tärkeimpänä tavoitteena on kehittää Suomeen Clarin-yhteensopiva tekninen infrastruktuuri ja kartuttaa sen puitteissa tarjottavia aineistoresursseja. Useiden suomalaisten yliopistojen lisäksi konsortioon kuuluvat Tieteen tietotekniikan keskus (CSC) ja Kotimaisten kielten tutkimuskeskus (Kotus) ovat avainasemassa teknisen infrastruktuurin tarjoajina. Niiden palveluiden piiriin on konsortion päämäärän mukaan tarkoitus koota kaikki olemassa olevat ja karttuvat sähköisesti käytettävät kieliaineistot ja -työkalut erilaisista yksittäisistä arkistoista.

Kotus on viime vuosina panostanut aineistojensa käytettävyyteen esimerkiksi vuonna 2006 avatulla Kaino-aineistopalvelullaan, jonka www-pohjaisen käyttöliittymän kautta aineistot on pystytty tarjoamaan helposti ja nopeasti entistä suuremmalle käyttäjäkunnalle (Suutari 2010b). Samansuuntaista työtä tehdään myös CSC:n palvelimella sijaitsevaa Kielipankkia kehittävissä Helsingin yli-

opiston Fin-Clarin- ja Meta-Nord-hankkeissa, joiden yhtenä tavoitteena on saattaa nykyistä laajempi valikoima Kielipankin aineistoja myös Unix-käyttöön tottumattomien tutkijoiden ulottuville. Nykyisin monet suomen ja muiden uralilaisten kielten erikokoisia korpuksia sisältävät kokoelmat ovat käyttöluvanhaltijoidensa hyödynnettävissä ainoastaan komentorivin käyttöosaamista vaativalla Kielipankin Unix-palvelimella, mutta tulevaisuudessa niistä voitaneen tehdä hakuja erilaisiin tarpeisiin verkon kautta toimivalla, käyttäjäänsä eteenpäin ohjeistavalla korpushakuohjelmalla.

Aineistojen monipuolisemman käytettävyyden parantamisen ohella hanke työskentelee Kielipankin aineistojen käyttöluokitusten selkiyttämiseksi ja yhdenmukaistamiseksi. Aineistot voidaan ehdotetun yhteiseurooppalaisen standardin mukaisesti luokitella kokonaan julkisiksi tai tutkimuskäyttöön tarkoitetuiksi, tai niiden käyttöoikeus voidaan rajoittaa luvanvaraiseksi (Oksanen, Lindén & Westerlund 2010). Tämän lisäksi CSC ottaa parhaillaan käyttöön uutta sähköistä käyttölupajärjestelmää Kielipankille. Aineistojen helppo käytettävyys, selkeä aineisto- ja käyttöoikeusluokitus ja käyttölupien sujuva hallinta kuuluvatkin pitkäjänteisen ja turvallisen säilytyksen ohella tärkeimpiin arkistoilta vaadittaviin palveluihin aineistojen yhteiskäytön sujuvoittamiseksi.

Aineiston tuottajan velvollisuudet ja hyödyt

Aineistojen yhteiskäyttöön ja jakamiseen liittyy kiinteästi sekä tekijänoikeuden että yksityisyydensuojan näkökulma. Kieliaineiston tekijänoikeus on pääsääntöisesti sen tuottajalla tai kerääjällä, ellei oikeutta ole luovutettu erikseen aineiston julkaisijalle tai muulle taholle. Aineiston osien, analyysien ja käännösten tekijänoikeus saattaa määrittyä erikseen kunkin osion laatijalle, mikäli aineiston käsittelyyn osallistuu useampi henkilö. Tekijänoikeus tuo tutkijalle mukanaan velvollisuuden aineiston takana olevia tahoja, haastateltuja tai muita oikeudenomistajia sekä aineistoa myöhemmin käyttäviä tutkimus- ja muita yhteisöjä kohtaan. Aineiston tuottamiseen osallistuneiden yksityisyydensuojasta ja tekijänoikeuden kunnioittamisesta huolehtimisen lisäksi hyviin tutkimuseettisiin periaatteisiin kuuluu aineiston myöhemmän käytön määritteleminen mahdollisia eri käyttötarkoituksia ajatellen. Tutkijan apuna myöhemmän käytön huomioon ottamisessa toimivat arkistojen tarjoamat valmiit, kansainvälisesti yhtenäiset käyttöoikeusluokitukset. Aineiston ja sen osien oikeudenomistajien vastuulle jää silti huolehtiminen myöhemmän käytön ja julkisuusasteen sopimisesta yhdessä informanttien ja muiden aineiston tuottamiseen osallistuneiden tahojen kanssa.

Aineiston tuottajana tutkijan tehtäväksi jää lisäksi huolehtia aineiston ja sen metatietojen työstämisestä arkiston ohjeistuksen mukaiseen muotoon siirrettävyyden, luokittelun ja käytettävyyden varmistamiseksi. Eri työvaiheissa käsiteltävän aineiston tulee taipua eri formaatteihin, ja esimerkiksi puheaineiston tie nauhurilta litteroiduksi ja mahdollisesti myös tarkempia analyysitasoja sisältäväksi aineistoksi kulkee monen välineen ja sovelluksen kautta. Valmis aineisto tulee samalla tavoin voida tarjota käyttäjille erilaisia tutkimustarpeita ajatellen ja mahdollisesti useamman eri ohjelman kautta käytettäväksi. Fin-Clarin-hanke on antanut Kielipankkiin tallennettavien

uusien aineistojen tuottajille ohjeeksi aineistojen toimittamisen xml-muodossa. Äänitallenteiden olisi hyvä lisäksi olla litteroituja ja aikakoodattuja siten, että litteraatio on automaattisesti yhdistetty äänitallenteen vastaavaan kohtaan, mikä helpottaa huomattavasti hakujen tekemistä aineistosta analysointivaiheessa. Puheaineiston litteroinnin lisäksi ei vaadita muita nimikointitasoja, mutta ihannetapauksessa toki aineistoa eri tutkimustarkoituksiin käyttäneet tutkijat oheistavat aineistoon omat analyysitasonsa ja jakavat ne edelleen uuteen tutkimuskäyttöön. Sen sijaan tiedot aineiston sisällöstä, metatiedot, laaditaan aineiston oheen kansainvälisten standardien ja arkiston ohjeiden mukaisesti. (Ks. esim. CLARIN Metadata Now 2009.)

Myöhemmän tutkimuskäytön mahdollistaminen on aineiston tuottajalle prosessi, johon aineiston varsinaisen keräämisen lisäksi kuluu aikaa ja vaivaa. Aineiston tuottaminen kestävään, myös muuta kuin omaa käsillä oleva tutkimusta palvelevaan muotoon, vaatii erityistä työpanosta, jonka tulee näkyä myös tutkijan tieteellisenä ansiona. Asianmukaisesti toimitettu raaka-aineisto voi palvella myöhempää tutkimusta ja tieteenalaa omalta osaltaan yhtäläisesti kuin sen keräämisen motivaationa oleva ja tutkijan omana analyysinä syntyvä julkaisu. Myöhempää tutkimusta palveleva tiedeyhteisön käytettäväksi toimitettu ja julkaistu aineisto tulee näistä syistä nähdä tutkijalle ansioksi luettavana erillisenä tieteellisenä tuotoksena esitelmien, artikkeleiden, ohjelmistojen ja patenttien joukossa. Julkaisuksi luettavalle aineistolle asetettavat vaatimukset määrittää lopulta tiedeyhteisö, joka määrittelemällä aineiston tieteelliset laatuvaatimukset nostaa samalla aineiston tuottajan työn arvoa ja tekee sen näkyväksi. Raaka-aineiston tai sen osan tuottajan kannalta sen jakaminen ja käyttöoikeuksin rajattu tai vapaa julkaiseminen merkitsee yhä uusia viittauksia aineistoon, joka pystyy todennäköisesti palvelemaan hyvinkin erityyppisiä tutkimuksia tuottajansa oman tutkimusintressin lisäksi. Aineiston arvostaminen tieteellisenä julkaisuna motivoi sen tutkijaa käsittelemään aineistonsa myös kansainvälisen käyttäjäkunnan huomioon ottaen.

Tavoitteena aineistonhallinnan yhteiset käytänteet

Aineistoa keräävän tutkijan kannalta suuri kysymys on, miten aineisto työstetään julkaisuksi arkistointia ja myöhempää käyttöä varten joustavasti ja omaa tutkimustyötä samalla tukien. Tutkijan on esimerkiksi ratkaistava, miten aineisto työstetään tiettyyn formaattiin, millä sovelluksilla litteroidaan, annotoidaan, kirjoitetaan käännökset ja käsitellään sanastoa, mitä aineiston käytöstä sovitaan informanttien kanssa ja miten voi selvittää eri lähteistä poimitun aineistokokonaisuuden tekijänoikeudet. Suurin osa kysymyksistä vaatii selvittämistä pelkästään aineiston kerääjän omia tutkimuskysymyksiä ja -julkaisuja ajatellen, mutta niiden ratkaisutapa riippuu samalla aineiston myöhemmästä arkistointi- ja käyttötavasta. Aineiston käsittelyyn ja tuottamiseen liittyviin kysymyksiin tarvitaankin yhteistä keskustelua ja aineistonhallinnan hyvien käytänteiden jakamista. Kokemuksen ja tutkimustapojen ja -tekniikoiden kehittymisen myötä päivitettävät aineiston työstämistä koskevat ohjeet ja suositukset tukevat yksittäisiä tutkijoita aineiston hallinnassa, ja ne kuuluvat teknisen raken-

teen lisäksi erottamattomasti osaksi toimivaa aineistoinfrastruktuuria.

Arkistot ja aineistoihin keskittyvät verkostot kuten Clarin voivat tarjota aineiston työstämiseen erilaisia suuntaviivoja kuten metatietostandardeja ja käyttöoikeusluokituksia. Suomessa FinClarin-konsortion perustaminen ja sen kautta tehtävä työ eri aineistokokonaisuuksien kokoamiseksi yhteen on hyvä alku. Konkreettiset aineiston työstämiseen liittyvät toimet ja käytänteet on kuitenkin valittava laajassa lingvistiyhteisössä. Aineistojen tärkeimmät tuottajat, tutkijat, ja aineistojen käyttäjinä tutkijoiden ohella myös opiskelijat, tietävät parhaiten, millaisia aineistosisältöjä ja -muotoja tarvitaan ja voidaan käyttää. Myös aineiston käsittelyyn liittyvät käytännön kysymykset, kuten millainen aineisto kannattaa litteroida Elan- tai Transana-ohjelmalla tai milloin informanttien kanssa tulisi käyttää kirjallista sopimuspohjaa, jäävät tutkijoiden ratkaistavaksi. Samankaltaisia asioita ei kannata tehdä eri yliopistoissa ja tutkimusryhmissä eri tavoin, mikäli siihen ei ole tutkimuksen tavoitteiden määrittelemää sisällöllistä tarvetta, vaan kokemuksia on syytä kerätä ja jakaa tutkimusyhteisössä entistä aktiivisemmin. Toisaalta kieliaineistoja tallentava arkistotahokaan ei saa jäädä pelkäksi aineistoja vastaanottavaksi tekniseksi rakenteeksi, vaan sen täytyy pystyä jatkuvaan vuoropuheluun aineistoa tuottavan ja käyttävän lingvistiyhteisön kanssa. Tutkijat ovat parhaita asiantuntijoita kertomaan aineistoihin liittyvistä käytännön haasteista ja tarpeista; arkisto puolestaan pystyy luomaan ja kehittämään erilaisten aineistojen hallintaan ja käyttöön sovellettavia kansainvälisen infrastruktuurin mukaisia teknisiä ja sisällöllisiä ohjeita ja sovelluksia.

Aineistoja on yliopistoissa ja tutkimuslaitoksissa kartutettu pitkäjänteisesti erilaisissa tutkimusprojekteissa ja myös opiskelijavoimin (puheaineistojen keruuhankkeista ks. Lyytikäinen & Yli-Paavola 2010; Siiroinen 2010; Karttunen & Rouhikoski 2010). Erilaisten hankkeiden yhteistyön ja suunnitelmallisuuden lisääminen eri tutkimusten tarpeita palvelevien aineistojen keruussa ja työstämisessä korostuukin varmasti tulevaisuudessa. Aineistonhallinnan kysymykset liittynevät tulevaisuudessa yhä kiinteämmin myös opetukseen, ja aineistoja ja niiden analyysejä voi olla järkevää kartuttaa systemaattisemmin myös kurssitöinä. Esimerkkinä tästä on Helsingin yliopiston suomen kielen ja suomalais-ugrilaisten kielten oppiaineryhmien suunnitelma kytkeä aineistojen käsittelyyn tarkoitettujen sovellusten käyttö ja olemassa olevien aineistojen analyysitasojen kartuttaminen yhä tiiviimmin osaksi opinnäyteseminaareja, pääaineeseen kytkettyjä tietoja viestintätekniikan opintoja, sukukielten kursseja ja kenttätyöopetusta.

Aineistonhallinnan käytänteiden luominen, kehittäminen ja vakiinnuttaminen, erilaiset aineistotalkoot ja johdonmukainen aineistonhallinta säästänevät tulevilta lingvistisukupolvilta paljon työtä, jota muuten tehtäisiin nykyaineistojen formaattien muokkauksen ja niiden käyttöoikeuksien selvittämisen parissa – elleivät nykyaineistot unohdu ja jää hyödyntämättä uudelleen. Ylimääräistä vaivaa on syytä välttää myös aineiston käsittelemisessä arkistojulkaisuksi, eikä kynnys tuottaa myös myöhempään käyttöön sopivaa aineistoa saa nousta liian korkeaksi tutkijalle. Kysymystä siitä, miten mahdollisimman vähällä vaivalla voitaisiin tuottaa mahdollisimman monikäyttöisiä aineistoja, on pidettävä jatkuvasti

esillä aineistoja käyttävän ja tuottavan lingvistiyhteisön ja tutkimushakkeiden piirissä, kieliarkistojen ja niiden kehittäjien työssä sekä näiden yhteisillä olemassa olevilla ja aktiivisesti rakennettavilla uusilla foorumeilla.

Lotta Jalava
etunimi.sukunimi@helsinki.fi

Lähteet

CLARIN Metadata Now. Short Guide, March 2009. http://www.clarin.eu/system/files/Metadata_now-CLARIN-ShortGuide.pdf. (30.12.2010.)

Grünthal, Riho 2010: Matkueita ja yksittäisiä tutkijoita. Suomalais-Ugrilaisen Seuran keruuretkien tausta ja tavoitteet. – *Uralica Helsingiensia* 4 s. 17–51.

Karttunen, Miia – Rouhikoski, Anu 2010: Kentällä kokeillen. Murresyntaksin tutkijat uusia aineistoja kokoamassa. – *Virittäjä* 114 s. 426–432.

Lyytikäinen, Erkki – Yli-Paavola, Jaakko 2010: Suomen kielen nauhoitearkisto 50-vuotias. – *Virittäjä* 114 s. 411–419.

Oksanen, Ville – Lindén, Krister–Westerlund, Hanna 2010: Laundry symbols and license management. Practical considerations for the distribution of LRs based on experiences from CLARIN. – *Proceedings of LREC 2010: Workshop on Language Resources: From Storyboard to Sustainability and LR Lifecycle Management. May 2010, Malta.* http://www.lrec-conf.org/proceedings/lrec2010/workshops/W20.pdf.

Saarinen, Sirkka 2007: Fennougristinen kenttätyö. http://www.kotus.fi/index.phtml?s=734. (30.12.2010.)

Siiroinen, Mari 2010: Aikamatkoja ja nykypäivää äänitallenteiden maailmassa. Suomen kielen nauhoitearkiston 50-vuotisjuhlavuoden satoa. – *Virittäjä* 114 s. 420–422.

Suutari, Toni 2010a: Suunvuoro. – *Virittäjä* 114 s. 323.

—— 2010b: Suomen kielen nauhoitearkisto – vireä viisikymppinen. – *Virittäjä* 114 s. 423–426.