Kielitieteellisen aineiston kerääjän opas

Niko Partanen

2018-11-20

Contents

Chapter 1

Esittely

Kielitieteellisiä aineistoja on kerätty jo satojen vuosien ajan, joten myös uudet aineistot on syytä nähdä tällaisessa syvemmässä historiallisessa kontekstissa. Suomalais-ugrilaisten kielten osalta voi sanoa, että 1800-luvun lopulla kerätyt aineistot alkavat transkriptioltaan ja kuvailutiedoiltaan muistuttaa 1900-luvun aikana kerättyjä. Tätä vanhemmat aineistot ovat myös hyvin arvokkaita, mutta niiden käyttöön liittyy eri tavalla tehtävä tulkinta. Selkeä ero on esimerkiksi siinä, että vanhimmat aineistot eivät useinkaan sisällä tietoja kielenoppaista. Kieltä ajateltiin ehkä eri tavalla puhujasta irrallisena entiteettinä. Toisaalta 1800-luvun lopun aineistoissa on usein yhtä paljon, tai vähän, kuvailutietoja kuin uudemmissakin aineistoissa.

1.1 Kuka minä olen?

Työskentelen Kotimaisten kielten keskuksessa kirjasto- ja aineistoyksikössä erityisasiantuntijana. Olen työskennellyt aiemmin mm. Freiburgin yliopistossa, Hamburger Zentrum für Sprachkorpora -keskuksessa sekä Pariisin ENS-yliopiston Lattice-laboratoriossa. Pääasiallinen kieli jota tutkin on komisyrjääni. Olen myös kiinnostunut itämerensuomalaisista kielistä ja permiläisistä kielistä yleisesti. Olen käyttänyt paljon kieliteknologiaa ja jonkin verran koneoppimismenetelmiä, mutta yleensä käyttäjän perspektiivistä, haluten saada näillä jonkin vaikean tai työlään työvaiheen pois käsistäni.

Teen väitöskirjaani komisyrjäänin murteellisesta variaatiosta. Olen julkaissut viime vuosina erilaisia artikkeleja, mutten ole varma, tuleeko näistä mikään väitöskirjaani. En ole varsinaisesti tehnyt asioita suositellun akateemisen urakehityksen mukaisesti. CV:ni on luettavissa mm. täällä.

1.1.1 Muutamia termejä avattuina

1. Korpus: Jotenkin valittu aineisto, johon on usein tehty erilaisia annotointeja
2. Annotointi: Jonkin asian systemaattinen merkintä, esim. sanaluokka, lauseenjäsen, elollisuus yms.
3. Lisenssi: Ehdot, joilla aineistoa saa käyttää ja jakaa

Chapter 2

Tutkimusetiikka

Tänne jotain.

2.1 Aineistojen käyttö

Periaatteessa mitä tahansa tietyllä kielellä olevaa materiaalia voi käyttää tutkimusaineistona. Toisaalta aiheen tarkempi miettiminen on hyvin kannattavaa. Usein haluat esimerkkejä jostain tietystä ilmiöstä tai asiasta. Lopullisessa analyysissasi voi olla hyvin tärkeää, millaisista konteksteista esimerkit tulevat. Tekstilaji, aika, käännöksellisyys ja muut seikat vaikuttavat väistämättä. Voit tietysti lähteä liikkeelle siitä, että keräät jonkin ilmiön kaikki esimerkit, jotka vain voit löytää, mutta tämä on käytännössä mahdoton tavoite. Tuloksena on, että olet saanut tietystä rajallisesta aineistosta ne esimerkit, jotka olet sattunut löytämään.

Näin ollen tutkimasi asian esiintymisfrekvenssi vaikuttaa suoraan siihen, kuinka paljon aineistoa tarvitset.

Annotoitujen korpusten hyvä puoli on siinä, että niiden sisältö on useimmiten valittu jonkin prinsiipin mukaan. Ihannetilanteessa tämä periaate on kuvattu jossain ja pystytään myöhemminkin perustelemaan järkevästi. Huomaa, että aina näin ei ole. Erityisesti kielten dokumentaatioprojekteissa tehdyt korpukset ovat usein ns. opportunistisia, ja tavoitteena on ollut kerätä kaikki mahdollinen. Se, mitä on lopulta litteroitu ja annotoitu, on kuitenkin useimmiten ollut hyvin tietoinen prosessi.

Myös annotaatiot on aina tehty jonkin tietyn periaatteen mukaan. Tavoite on aina, että samaa annotaatiomallia käytetään systemaattisesti koko aineistossa. Tässä tavoitteessa harvoin onnistutaan, ja jokaiseen korpukseen jää jonkinlaisia virheitä. Älä luota yhdenkään aineiston olevan täydellisen systemaattinen. Jos jokin asia ei ole kunnossa, voit myös aina korjata sen itse. Voitko jakaa korjauksesi muiden kanssa riippuu mm. lisensseistä.

2.2 Lisenssit ja tekijänoikeudet

Kaikkea aineistoa ei ole lisensoitu. Lisenssien puuttumattomuus on erityisen yleistä tietyillä kulttuurialueilla, esimerkiksi Venäjällä. Lisenssi käytännössä tarkoittaa sitä, että aineiston tekijä antaa muille oikeuden tehdä aineistolla tiettyjä asioita.

Lisensoimatonta materiaalia ei periaatteessa voi käyttää kuin tietyin rajallisin ehdoin.

Avoimet lisenssit ovat erityisen pitkällä ohjelmistokehityksen puolella.

Kielitieteellisiin aineistoihin käytetään usein lisenssiä CC-BY. Tämä tarkoittaa, että aineiston käyttäjän on viitattava alkuperäiseen aineistoon ja tekijään. Aineistoista voi tehdä myös uusia versioita, mutta niidenkin yhteydessä on aina oltava viittaus.

Viittaus on pakko tehdä tieteellisessä käytössä myös hyvän tieteellisen käytännön mukaisesti. Ei ole selvää, kuinka tiukasti CC-lisenssien rikkomisia valvotaan, ja käytännössä lisenssi on aina peli siitä, kenellä on eniten rahaa. Suuret yritykset rikkovat tekijänoikeuksia säännöllisesti sen turvin, ettei loukatuilla tahoilla ole resursseja haastaa heitä oikeuteen. Toisaalta hyvän tieteellisen käytännön rikkomisesta on melko selviä toimintamalleja ja rangaistuksia.

2.2.1 Lisenssityypit

- CC-BY tarkoittaa, että lähteeseen on viitattava
- CC-BY-NC kieltää kaupallisen käytön
  - On erittäin vaikeaa määritellä, mitä on kaupallinen käyttö
- CC-BY-ND kieltää johdannaisteokset (no derivations)
  - Tiukasti tulkittuna mikä tahansa jatkokäyttö menee tämän allen
- ACA ei varsinaisesti ole lisenssi, vaan käyttöehto, että aineistoa saa käyttää tutkimustyössä. Sitä ei siis saa jakaa edelleen, ja käytännössä kaikki käyttörajoitukset ovat voimassa.
- CC-0 / Public Domain ei ole lisenssi, vaan oikeastaan lisenssin puuttuminen

CC-0 on ajoittain esitetty suositelluksi lisenssiksi. On hieman epäselvää, salliiko laki kuitenkaan tekijänoikeuksista luopumista. Toisaalta ei ole selvää, tajuavatko tutkijat, että jo tieteellinen käytäntö pakottaa viittaamaan.

2.2.2 Public Domain

On tiettyjä tapauksia, jolloin materiaali muuttuu tekijänoikeusvapaaksi. Selkein tapaus on ikä. 70 vuotta tekijän kuolemasta on Suomessa vallitseva raja to-menetykselle. Esimerkiksi Venäjällä tilanne voi olla erilainen, sillä esimerkiksi kansallissankareille ja sodan vainojen uhreille on erilaisia säädöksiä tämän suhteen.

Tekijänoikeus voi myös raueta, jos kyseessä on orpoteos. Tämä tarkoittaa, että tekijällä ei ole perillisiä, joilla olisi oikeus teokseen. Tällöin hyvinkin tuore teos voisi periaatteessa olla lisenssivapaa. Käytännössä tämän todistaminen Suomessa on ilmeisen hankalaa, mutta esimerkiksi Venäjän kirjastot ovat tietyissä tapauksissa tehneet tätä.

Tietyn teoksen olettaminen tekijänoikeusvapaaksi on aina tietty riski. Turvallisinta se on silloin, kun jokin kansallisesti virallinen taho on jo tehnyt määritelmän puolestasi. Tällöinkin viittaaminen juuri tuohon versioon on käyttäjän kannalta selkeintä ja turvallisinta.

2.2.3 Tutkimusaineistojen koonti ja lisenssi

Kun keräät tutkimukseesi materiaalia eri lähteistä, teet käytännössä uutta aineistoa. Sinun on otettava käyttämiesi materiaalien lisenssit, jotta voit tietää, saako aineistoasi jakaa eteenpäin ja millaisin ehdoin. Nykyisin on modernia ja toivottavaa esimerkiksi jakaa tutkimusaineistosi artikkelisi tai gradusi liitteenä. Tai voit rekisteröidä käyttämäsi aineiston esimerkiksi Zenodo-palvelussa, ja viitata siihen erikseen. Näin jokainen julkaisu olisi käytännössä eräänlainen kaksoisjulkaisu, jossa julkaisisit erillisesti tutkimuksesi ja siinä käyttämäsi aineiston.

Jos tästä halutaan tehdä oikein fiiniä, olisi syytä julkaista myös erillinen teksti, jossa kuvaat käyttämäsi aineiston. Suuremman tutkimuksen kuten gradusi yhteydessä tuo voisi hyvin olla jollain tavalla osa metodilukua.

2.3 Annotaatioiden välttämättömyys ja riittämättömyys

Kuten totesin, annotoidut aineistot ovat usein hyvin hyödyllisiä. Tehokas hakeminen niistä mahdollistaa monessa kohtaa nopeamman halutun otoksen saamisen, jopa sen täydellisen automatisoinnin. Tämä ei

kuitenkaan tarkoita, että työ olisi heti tehty. Tutkimuskysymyksestä riippuen esimerkkien syvällisempi annotointi on usein välttämätöntä. Tämä sinun on tehtävä itse.

On myös hyvin mahdollista, että olet eri mieltä korpuksen annotointien kanssa. Mieti tällöin, koodaavatko annotaatiot saman tiedon, kuin mitä itsekin merkitsisit, vai onko ongelma siinä, että jotain distinktiota oikeasti ei merkitä. Jos et pidä valitusta mallista, on se mahdollista aina muuttaa johonkin muuhun formaattiin, mutta puuttuvaa tietoa ei saa sinne kuin itse lisäämällä.

Hyvin normaali tilanne on, että kun teet jonkinlaisen haun korpukseen, sisältää se esimerkiksi 95% juuri sitä mitä haluat. Joukossa on joka tapauksessa lähes aina jotain turhaakin, esimerkiksi annotaatiovirheitä tai johonkin muuhun liittyviä esimerkkejä. Joka tapauksessa näiden käyminen läpi jatkoannotaatiotyön yhteydessä on verrattain nopeaa. Mieti myös tarkkaan sitä, onko korpuksessa nykyisen haun ulkopuolelle jääviä esimerkkejä, jotka kuitenkin tarvitsisit.

Yksi hyvä formaatti aineiston käsittelylle on perinteinen taulukkolaskentaohjelma, kuten LibreOffice tai Excel. Excelin kanssa kannattaa olla tarkkana, että se ei sotke fontteja. Yksi hyvä vaihtoehto on niin sanottu CSV-tiedosto. Tämä tarkoittaa comma-separated-values. Pilkun sijasta erottava merkki voi myös olla sarkain tai puolipiste. Tällaisen voi myös avata Excelissä ja muualla, ja se on monissa tapauksissa hyvin elegantti vaihtoehto.

2.4 Aineiston ja tutkimuksen erillisyys ja yhteys

2.5 Versiokontrolli

Chapter 3

Erilaiset aineistot

Mielestäni tutkijan kannattaisi pyrkiä siihen, että hänen käyttämänsä aineisto olisi digitaalisessa formaatissa. Tietysti on mahdollista etsiä esimerkkejä myös käymällä kirjoja läpi omin silmin, mutta nykyään on ehkä usein muitakin vaihtoehtoja.

Tekstitiedostot käyvät sinällään jo moneen. On syytä tutustua säännöllisiin lausekkeisiin, tai ainakin erilaisiin hauissa toimiviin wild card -merkkeihin. Ohjelmat kuten AntConc ovat kokeilemisen arvoisia.

Jos haluat käyttää aineistona jotain kirjaa tai muuta teosta, joka sinulla on esimerkiksi skannattuna, saa tekstin siitä nykyään helposti ulos OCR-ohjelmilla. Kerron mielelläni lisää!

Iso ero on myös puhuttujen ja kirjoitettujen aineistojen välillä. Puhutut aineistot tulevat usein formaateissa, jotka mahdollistavat lisäannotointien teon. Esimerkiksi ELAN ja Exmaralda-ohjelmien tuottamat tiedostot ovat sellaisia, että niissä on omat kerrokset erilaisilla annotointityypeille. Voit siis tehdä uuden kerroksen, jossa on annotoituna vain sinun tutkimuksellesi tarpeelliset asiat niissä esimerkeissä, joita haluat käyttää.

3.1 Aineistoesimerkkejä

- Kielipankki
- Oslon yliopiston korpukset
  - Esimerkiksi Ruija-korpus
  - LIA sápmi hyvä esimerkki akateemisesta lisenssistä
- HZSK
  - Selkuppi, nganasaani
- Universal Dependencies
- Ob-Babel
- Oulun saameaineistot

3.1.1 Venäjällä tyypillisiä aineistoja

- Komin kansallinen korpus
- Udmurt corpus
- Beserman corpus
- Selkup.org
- **Ongelma usein puutteellinen export-toiminto, mutta asiat ehkä kehittyvät.**

3.1.2 Satunnaista

- Tundra Nenets sample sentence corpus
- Valentin Gusevin nganasaani-korpus

Chapter 4

Tutkimusetiikka

> Tieteellisen tutkimuksen luotettavuus ja tulosten uskottavuus edellyttävät, että tutkimuksessa noudatetaan hyvää tieteellistä käytäntöä. Vastuu hyvän tieteellisen käytännön noudattamisesta kuuluu koko tiedeyhteisölle ja jokaiselle tutkijalle. Helsingin yliopisto on sitoutunut Tutkimuseettisen neuvottelukunnan ohjeisiin hyvästä tieteellisestä käytännöstä ja sen loukkausten käsittelemisestä. HY

Kielitieteelliset aineistot poikkeuksetta käsittelevät ihmisiä. Joku on tuottanut tai kirjoittanut lauseet, joita tutkimme.

4.1 Tietosuojalain nykytila

> Eduskunnan on tarkoitus hyväksyä asetusta täydentävä tietosuojalaki syksyllä 2018. Asetuksen edellyttämät kansalliset lainsäädäntömuutokset ja yksityiskohdat ovat vielä auki. Tietosuoja-asetusta on tästä huolimatta sovellettava 25.5.2018 alkaen. (lähde)
>
> Kansallista liikkumavaraa yleisen tietosuoja-asetuksen täydentämiseksi ehdotetaan käytettävän tilanteissa, joissa henkilötietolain kumoamisesta seuraisi tarve kansalliseen sääntelyyn, esimerkiksi henkilötietojen käsittelyn oikeusperusteen säilyttämiseksi eräissä tilanteissa tai tieteellisen tutkimuksen edellytysten säilyttämiseksi mahdollisimman pitkälle nykyisen kaltaisina. (lähde)
>
> Edellytyksenä on niin ikään, että henkilörekisteriä käytetään ja siitä luovutetaan henkilötietoja vain historiallista tai tieteellistä tutkimusta varten sekä muutoinkin toimitaan niin, että tiettyä henkilöä koskevat tiedot eivät paljastu ulkopuolisille, ja että henkilörekisteri hävitetään tai siirretään arkistoitavaksi tai sen tiedot muutetaan sellaiseen muotoon, ettei tiedon kohde ole niistä tunnistettavissa, kun henkilötiedot eivät enää ole tarpeen tutkimuksen suorittamiseksi tai sen tulosten asianmukaisuuden varmistamiseksi. (lähde)

4.2 GDPR

Kesällä 2018 EU:n yleinen tietosuoja-asetus (General Data Protection Regulation, GDPR) astui voimaan. Se antaa yksityishenkilöille paremman suojan heidän henkilötiedoilleen ja keinot hallita niiden käsittelyä.

Tutkija, joka kerää tutkimusaineistoa, käsittelee lähes poikkeuksetta henkilötietoja. On mahdollista kerätä tietoja esimerkiksi anonyymin lomakkeen kautta, mutta tutkimuksen itsensä kannalta on yleensä tärkeä tietää kuka puhuja on, mikä on hänen sukupuolensa, ikänsä, syntymäpaikkansa tai äidinkielensä. Tarkka kerättävä tieto riippuu tutkimuskysymyksestä.

Henkilötietoja ovat esimerkiksi [(lähde: Tietosuoja.fi)(https://tietosuoja.fi/gdpr)]:

- nimi
- kotiosoite
- sähköpostiosoite, kuten etunimi.sukunimi@yritys.com
- puhelinnumero
- henkilökortin numero
- auton rekisterinumero
- paikannustiedot
- IP-osoite
- potilastiedot
- isovanhempien perinnöllisiä sairauksia koskevat tiedot

Näiden ohella on tässä yhteydessä mainittava, että myös **ääni** on henkilötieto. Tämä johtuu yksiselitteisesti siitä, että ihmiset pystyvät erittäin tarkasti tunnistamaan kenen tahansa tutun henkilön äänen. Toisaalta puheen automaattisen tunnistamisen, johon liittyy puhujan tunnistus, tekniikat kehittyvät niin nopeasti, että jos äänitiedosto on kuunneltavissa, on tämä käytännössä sama kuin että ainakin nimi olisi myös saatavilla. Tämä täytyy ottaa huomioon.

Henkilötietojen käsittelyyn tarvitaan peruste. Tällaisia ovat esimerkiksi:

- rekisteröidyn suostumus
- sopimus
- rekisterinpitäjän lakisääteinen velvoite
- elintärkeiden etujen suojaaminen
- yleinen etu ja julkinen valta
- rekisterinpitäjän tai kolmannen osapuolen oikeutettu etu.

Ongelma suostumuksen ja sopimuksen käytössä käsittelyperiaatteena on se, että ne mahdollistavat käytön vain tiettyyn tai useampaan sopimuksessa määriteltyyn tarkoitukseen. Ei ole laillisesti mahdollista sopia, että kerättyä aineistoa käytetään yleisesti kielitieteellisessä tutkimuksessa tulevaisuudessa.

Perinteisesti kielitieteellisen aineistonkeruun yksi päätavoite on kuitenkin ollut tallentaa näytteitä katoavista ja uhanalaistuvista kielimuodoista tulevaa käyttöä varten. Tämä vastaa hengeltään hyvin yleisen edun määrittelyä laissa:

> Yleisen edun mukaista käsittelyä voi esimerkiksi olla henkilötietojen käsittely tieteellisen tai historiallisen tutkimuksen tai tilastoinnin tarkoituksia varten.

Näin ollen akateemiseen instituution affilioituneen tutkijan tai projektin työ melko yksiselitteisesti menisi tämän alle.

Aineistojen pitkäaikaissäilytys ja muut arkistointikysymykset selkenevät vasta kun Suomen uusi arkistolaki valmistuu. Käytännössä kuitenkin normaali tilanne olisi, että tutkijat voisivat tutkimusprojektiensa yhteydessä arkistoida aineistot esimerkiksi Kotimaisten kielten keskukseen ja Kielipankkiin, mistä he itsekin saisivat ne uudelleen mahdollisia uusia tutkimuksiaan varten.

Käytännössä tapana on ollut, että tutkijat keräävät aineistoa ja säilyttävät sitä omilla tietokoneillaan vuosikymmeniä. Käytännössä on kuitenkin lain ja etiikan kannalta täysin mahdotonta, että tutkijat, yksityishenkilöinä, säilyttävät henkilötietoja sisältävää aineistoa tällä tavalla.

4.3 Data management plan

Tietojen käsittelystä on tehtävä suunnitelma seuraavissa tapauksissa:

- henkilötietojen käsittely organisaatiossa ei ole satunnaista
- organisaation vastuulla oleva henkilötietojen käsittely todennäköisesti aiheuttaa riskin rekisteröidyn oikeuksille ja vapauksille tai

- organisaatiossa käsitellään arkaluonteisia tietoja.

Nähdäkseni tutkijoiden tekemä aineiston käsittely väistämättä menee kategoriaan, jossa henkilötietojen käsittely ei ole satunnaista. Näin ollen jonkinlainen suunnitelma olisi syytä tehdä. Käytännössä tällaisia vaaditaan nykyään enenevissä määrin myös projektihakemusten yhteydessä, joten tässä mielessä luvassa ei ole mitään uutta.

Kyseessä on organisaation sisäinen asiakirja, jonka perusteella tietojenkäsittelytoimien lainmukaisuutta voidaan arvioida. Nähdäkseni siinä olisi syytä kuvata, kuka aineistoa käsittelee, missä ympäristössä, minne sitä kopioidaan, ja kuinka aineisto järjestetään.

4.4 Erityiset henkilötiedot

Erityiset henkilötiedot koskettavat tutkimus siinä mielessä hyvin voimakkaasti, että **niiden käsittely on lähtökohtaisesti kiellettyä**. Erityisiksi henkilötiedoiksi lasketaan seuraavat:

- rotu tai etninen alkuperä
- poliittisia mielipiteitä
- uskonnollinen tai filosofinen vakaumus
- ammattiliiton jäsenyys
- terveyttä koskevia tietoja
- seksuaalinen suuntautuminen tai käyttäytyminen
- geneettisiä ja biometrisia tietoja henkilön tunnistamista varten

Jos voit tehdä tutkimuksesi keräämättä erityisiä henkilötietoja, niin toimi näin. Tällä hetkellä, ymmärtääkseni, monikaan arkisto ei suostu ottamaan tällaista aineistoa vastaan, eikä sitä saa käsitellä esimerkiksi CSC:n IDA-ympäristössä.

Joka tapauksessa tämäntapainen käsittely olisi silti ehkä mahdollista (lähde: Tietosuoja.fi):

> Kun käsittely on tarpeen **yleisen edun mukaista arkistointia, tieteellistä ja historiallista tutkimusta tai tilastointia varten** tietosuoja-asetuksen mukaisesti unionin oikeuden tai jäsenvaltion lainsäädännön nojalla. Käsittelyn mahdollistavan sääntelyn tulee olla oikeassa suhteessa käsittelyn tavoitteeseen nähden, ja siinä tulee noudattaa keskeisiltä osin oikeutta henkilötietojen suojaan. Tässä yhteydessä on myös säädettävä toimenpiteistä, joilla suojataan rekisteröidyn perusoikeudet ja edut.

Tässä yhteydessä odotamme siis käsittääkseni kansallista tietosuojalainsäädäntöä, joka toivottavasti selventäisi arkistoinnin ja tieteellisen käytön.

Chapter 5

Tiedon käsittely

5.1 Anonymisointi

5.2 Pseudonymisointi

5.3 IDA-ympäristö