2.2.1 Perusjoukko ja otos

Yleisesti ottaen tilastomenetelmiä käytetään, kun pyritään tekemään päätelmiä jostain perusjoukosta (engl. *population*), jota ei kuitenkaan voi mitata kattavasti. Tällaisia perusjoukkoja voivat olla vaikkapa:

a. maailman kielet

b. suomen kielen puhujat

c. arabian puhujat Suomessa

d. viron kielen transitiiviset verbit kielenkäytössä

e. kolmivuotiaan käyttämät substantiivit

Perusjoukko riippuu aina kyseisestä tutkimuksesta, joten perusjoukon määrittely on tutkimusprosessin keskeinen vaihe. Se määritellään tutkimuksen alussa, mutta siihen voi joutua palaamaan vielä pohdinnoissa. Perusjoukko on se joukko, josta tutkija haluaa tutkimuksensa avulla tehdä päätelmiä. Tutkimuksen edetessä voi kuitenkin joskus huomata, että oma aineisto ei riitä siihen, että sen perusteella voisi tehdä yleistyksiä haluamastaan perusjoukosta. Esim. jos oma aineisto koskee lasten kielen oppimista esikouluvaiheessa, ei välttämättä riitä se, että tutkii vain yhden päiväkodin esikoululaisia, jos haluaa tehdä yleistyksiä sen päiväkodin ulkopuolelle, esim. koskemaan koko kyseistä kaupunkia, maakuntaa, tai maata.

Perusjoukko muodostuu havaintoyksiköistä. Niitä kutsutaan myös tilastoyksiköiksi tai tutkimusyksiköiksi. Jos perusjoukko on maailman kielet, tällöin havaintoyksikkö on kieli. Jos perusjoukko on suomen kielen puhujat, tällöin havaintoyksikkö on yksittäinen suomen kielen puhuja. Jos perusjoukko on suomen kielen *olla*-verbin esiintymät, tällöin havaintoyksikkönä on suomen kielen *olla*-verbin yksittäinen esiintymä kielenkäytössä. Tutkimuksessa määritellään, mitä ominaisuuksia havaintoyksiköistä halutaan analysoida.

Analysoitavaa ominaisuutta kutsutaan muuttujaksi ja muuttuja voi saada analyysin perusteella erilaisia arvoja. Samasta havaintoyksiköstä analysoidut muuttujat muodostavat yhdessä havainnon ja aineiston kaikki havainnot muodostavat yhdessä havaintoaineiston. (ks. Luojola 2006: 17).

Esimerkiksi suomen *olla*-verbin esiintymistä voidaan analysoida niiden aikamuodot, esiintyminen kieltorakenteessa, järjestys suhteessa subjektiin, puhujan sukupuoli, puhujan ikä, jne. Nämä ovat *olla*-verbin joitain mahdollisia eri muuttujia ja niiden valinta ja relevanssi riippuu tietenkin tutkimuskysymyksestä. Muuttuja aikamuoto voi suomessa saada eri arvoja, kuten mennyt aika ja preesens; muuttuja sanajärjestys voi saada esim. arvoja subjekti–verbi-järjestys, verbi–subjekti-järjestys ja ei ilmi-subjektia; jne. Muuttujan tarkemmat arvot selviävät aina lopullisesti itse aineistosta. *Olla*-verbiä koskeva havaintoaineisto muodostuisi kaikista *olla*-verbin esiintymistä ja niistä analysoidut muuttujat arvoineen.

Havaintoaineisto järjestetään yleensä havaintomatriisin muotoon, jossa muuttujat esitetään sarakkeissa ja havaintoyksiköt rivillä. Tyypillisesti ensimmäinen sarake sisältää havaintoyksikön tunnuksen. Havaintoyksikön tunnus voi olla vaikkapa koehenkilön tunnus aineistossa, kielen tunniste ISO639.3-järjestelmässä tai sanan esiintymän tunnus korpuksessa (esim. järjestysluku). Muut sarakkeet sisältävät havaintoyksiköstä kerättyjen muuttujien arvot ja ehkä jotain metatietoa havaintoyksiköistä (kuten koehenkilön ikä, sukupuoli, asuinpaikkakunta jne.). Kullakin rivillä on yhdestä havaintoyksiköstä kerätyt tiedot.

Usein tutkimuksessa on mahdotonta kerätä tietoa koko perusjoukosta, koska se veisi aivan liikaa aikaa ja resursseja tai olisi muulla tavoin käytännössä mahdotonta. Tilastollisessa tutkimuksessa perusidea onkin, että perusjoukon käyttäytymistä pyritään arvioimaan ottamalla siitä otos. Keskeinen menetelmällinen kysymys tällöin on: **miten hyvin tulos on yleistettävissä otoksesta perusjoukkoon?**

Tuloksen yleistettävyys otoksesta perusjoukkoon riippuu etenkin siitä, miten hyvin otos edustaa perusjoukossa esiintyvää vaihtelua. Tällöin keskeisiä kysymyksiä ovat, miten edustava (engl. *representative*) otos on, sekä onko otos tasainen (engl. *unbiased*), eli että se ei ole jollain tavoin vinoutunut. Jos otos ei edusta perusjoukkoa kovin hyvin, sen perusteella ei voi yleistää otoksen ulkopuoliseen aineistoon kovin luotettavasti. Otoksen

muodostuksesta lisätietoa korpustutkimuksessa tarjoaa esimerkiksi Arppe (2008: 66–67) [https://helda.helsinki.fi/handle/10138/19274]; otannasta lapsen kielen oppimisen tutkimuksessa ks. Gries (2013: 22) [https://helsinki.primo.exlibrisgroup.com/discovery/fulldisplay?docid=alma9926962003506253&context=L&vid=358UOH_INST:VU1&lang=fi&search_scope=MyInst_and_CI&adaptor=Local%20Search%20Engine&tab=Everything&query=any,contains,statistics%20for%20Linguistics%20with%20R&offset=0].

Vinoumia aineistoon voi aiheuttaa mm. ne rajoitukset, mistä perusjoukon jäsenistä on edes saatavilla aineistoa. Sitä osaa perusjoukkoa, josta on ylipäätään saatavilla aineistoa, nimitetään joskus kehykseksi (engl. *frame*; Bell 1978: 126). Historiallisessa sosiolingvistiikassa kehys tarkoittaa esimerkiksi sitä historiallisten kirjeiden ja muiden kirjallisten tuotosten joukkoa, joka on säilynyt nykypäivään asti ja joka on toimitettu tutkijoiden käyttöön. Korpustutkimuksessa kehys voisi tarkoittaa niiden tekstilajien joukkoa, jotka ovat edustettuna tietyssä korpuksessa. Typologiassa kehys tarkoittaa sitä joukkoa maailman kielistä, joista on saatavilla riittävän luotettavaa tutkimusmateriaalia (esim. kielioppikuvauksia). Kokeellisessa tutkimuksessa kehys voisi tarkoittaa sitä henkilöiden joukkoa, josta koehenkilöitä on ylipäätään mahdollista saada mukaan tutkimukseen. Otos puolestaan on se kehyksen osajoukko, joka on otannalla valittu tutkimukseen.

Miten sitten laaditaan edustava ja tasainen otos? Yksinkertainen satunnaisotanta olisi klassisten tilastomenetelmien näkökulmasta kaikkein paras, mutta se voi olla käytännössä vaikea toteuttaa. Satunnaisotannan avulla havaintoyksiköt valitaan perusjoukosta satunnaisesti, mikä takaa sen, että perusjoukon jäsenillä on sama todennäköisyys tulla valituksi otokseen. Satunnaisuuden avulla otoksesta tulee edustava ja vinoumaton, koska havaintoyksiköiden valinta on toisistaan riippumaton eikä riipu tutkijan tekemistä valinnoista.

Satunnaisotannan lisäksi on käytössä erilaisia otantamenetelmiä. Suosittelen katsomaan oheisen lyhyen englanninkielisen videon, joka esittelee tyypillisiä otantamenetelmiä: https://www.youtube.com/watch?v=be9e-Q-jC-0

Videossa painotetaan mm. sitä, että populaation luonne ja käytettävissä olevat resurssit vaikuttavat pitkälti siihen, millaista otantaa kannattaa käyttää.

Kielitieteessä usein käytetty menetelmä on ns. ositettu satunnaisotanta, joka yhdistää kaksi erilaista otantamenetelmää. Ositetussa otannassa perusjoukko jaetaan osittaisiin, esim. kielet kielikuntiin. Tutkija valitsee ensin satunnaisesti ositteet (esim. riittävän määrän kielikuntia) ja sen jälkeen osittaisten sisältä varsinaiset havaintoyksiköt (esim. kielet). Osituksen avulla voidaan varmistaa, että otos edustaa populaatiota mahdollisimman hyvin ainakin ositukseen valittavalla tasolla. Kielitypologiassa tällaisia osittaisia ovat esim. kieliperheet ja maantieteelliset alueet. Korpustutkimuksessa osittaisia ovat esimerkiksi erilaiset tekstilajit tai kirjoittajat. Sosiolingvistiikassa osittaisia voivat olla esimerkiksi eri murteet; kielenopetuksessa esimerkiksi eri koulut.

Mikään otantamenetelmä ei ole täydellinen, vaan jokaiseen niihin liittyy ongelmia. Klassiset tilastomenetelmät edellyttävät, että otannassa on käytetty satunnaisotantaa; muutoin niitä menetelmiä ei periaatteessa voisi käyttää yleistämään otoksesta perusjoukkoon. Tästä vaatimuksesta ei tosin aina pidetä kiinni kovin tiukasti. Satunnaisotantaa käytetään, jotta havaintoyksiköiden valinta ei vinouttaisi aineistoa. Satunnaisotannan avulla jokaisella perusjoukon (tai tarkasti ottaen kehyksen) jäsenellä on sama todennäköisyys tulla valituksi otokseen ja näin ollen voidaan sanoa, että kunkin havaintoyksikön valinta on riippumaton muiden havaintoyksiköiden valinnasta. Jos näin ei ole, saatamme tuottaa systemaattisen vinouman otokseen ja päätyä vääriin johtopäätöksiin perusjoukon käyttäytymisestä.

Keskeinen ongelma satunnaisotannassa on, että aito satunnaisotanta on harvoin mahdollinen. Korpustutkimuksessa satunnaisotantaa on kritisoitu seuraavalla tavalla. Esimerkiksi, valitaan aineistosta jokin tutkittava sana, järjestetään sen esiintymät satunnaisesti ja valitaan tästä joukosta ensimmäiset 20 % esiintymistä. Tällainen satunnaisotanta tuottaa kyllä toisistaan riippumattomia virkkeitä, mutta nuo virkkeet ovat irrallaan diskurssista eivätkä muodosta koherenttia jaksoa. Arppe (2008: 50 ja viittaukset; [https://helda.helsinki.fi/handle/10138/19274]) esittää, että olisi parempi ottaa paljon esimerkkejä useilta eri kirjoittajilta, jolloin myös eri kirjoittajien omaperäisiä tapoja voidaan kontrolloida.

Kielitypologiassa ongelma liittyy isolaatteihin. Isolaatti on kieli, jolle ei ole voitu löytää yhtään sukulaiskieltä (esim. baski). Se muodostaa siis yksin oman kielikuntansa. Jopa kolmasosa maailman kieliperheistä on tällaisia kieliä (Campbell 2017). Isolaattien vuoksi

typologiassa on käytännössä mahdotonta tehdä esimerkiksi ositettua satunnaisotantaa kielikuntien sisällä (Janssen et al. 2006; https://doi.org/10.1515/LINGTY.2006.013).

Kielentutkimuksessa ei ole selvää standardia otantaan. Parasta opiskelijan ja tutkijan kannalta onkin selvittää, mitä viimeaikaisin kirjallisuus suosittelee otannasta kullakin kielitieteen alalla. Tämä ei ole missään mielessä ideaali tilanne ja toivottavaa olisikin, että selvempiä standardeja saataisiin kehitettyä tulevaisuudessa. Loppukaneettina voitaisiinkin sanoa, että otannassa on kyse tutkimusresurssien tehokkaasta käytöstä: pyri valitsemaan sellainen otantamenetelmä, joka antaa parhaan mahdollisen tuloksen niillä aika- ja raharesursseilla, joita sinulla on käytössä. Menetelmän valinnassa ei siis kannata olla idealistinen, koska se voi osoittautua resurssien tuhlaamiseksi.

Käsittelen videossa ”Tilastollinen otanta”, miten satunnaisotannan kanssa pääsee liikkeelle EXCEL:ssä. EXCEL-tiedosto löytyy tästä oppimateriaalista nimellä ”Otanta”.

Otannan yhteydessä nousee usein kysymys, mikä on tarpeeksi suuri **otoskoko** eli miten paljon aineistoa pitäisi analysoida. Nyrkkisääntönä voitaisiin sanoa, että mitä voimakkaampi suhde muuttujien välillä esiintyy, sitä pienemmällä otoksella se on havaittavissa. Toisaalta kääntäen voidaan todeta, että mitä heikommasta mutta silti todellisesta suhteesta on kyse, sitä suurempi otos tarvitaan sen tunnistamiseksi. Otoskoko siis vaikuttaa siihen, miten todennäköisesti aineistosta löytää tilastollisesti merkitseviä tuloksia. Erittäin suurista aineistoista löytyy lähes väistämättä tällaisia tuloksia, joten niiden kohdalla on syytä arvioida aina muuttujien välisen suhteen voimakkuutta tarkkaan. Jos käytät khiin neliö -testiä 2x2-taulukolle, jo 50–100 tapausta sisältävä aineisto on melko varmasti riittävä varsinkin gradussa. Väitöskirjassani analysoin kielitypologista aineistoa 50 kielestä ja se riitti havaitsemaan khiin neliö -testillä muuttujien välisen riippuvuuden, joka oli voimakas. Jos muuttujilla on useampia luokkia, tällöin aineistoa tarvitaan jonkin verran enemmän. Asiasta on parasta keskustella oman ohjaajan kanssa.

Huom! Riittävän suuri otoskoko voidaan määritellä eri testeille matemaattisesti **tilastollisen voiman** avulla. Tilastollisesta voimasta voit lukea lisää Tero Vahlbergin oheisesta esityksestä [http://www.orl.fi/wp-content/uploads/2017/06/Tilastotiede_Otoskoko.pdf]. Tilastollisen voiman laskenta onnistuu kyllä EXCEL:issä, mutta vaatii lisäohjelman lataamisen (esim. https://www.real-

statistics.com/sampling-distributions/statistical-power-sample/). R:ssä suosittelen käyttämään pakettia pwr [https://cran.r-project.org/web/packages/pwr/index.html].

Viitteet:

Arppe, Antti 2008. *Univariate, Bivariate, and Multivariate Methods in Corpus-Based Lexicography – a Study of Synonymy*. Väitöskirja, yleinen kielitiede, Helsingin yliopisto. https://helda.helsinki.fi/handle/10138/19274

Bell, Alan 1978. Language samples. Teoksessa Joseph Greenberg, Charles Ferguson & Edith Moravcsik (eds.), *Universals of Human Language, vol. 1: Method and Theory*, 123–156. Stanford: Stanford University Press.

Campbell, Lyle R. (ed.). 2017. *Language Isolates*. London: Routledge.

Gries, Stefan Th. 2013. *Statistics for Linguistics with R*, 2. korjattu ja laajennettu laitos. Berlin: De Gruyter Mouton.

Janssen, Dirk, Balthasar Bickel & Fernando Zúñiga 2006. Randomization tests in language typology. *Linguistic Typology* 10(3): 419–440. https://doi.org/10.1515/LINGTY.2006.013