Johdatus tilastolliseen kielentutkimukseen (Kaius Sinnemäki)

Transkriptio

Tervetuloa tähän metodipankin osioon ”Tilastollisten menetelmien perusteita kielentutkijalle”. Ennen kuin päästään varsinaisesti vauhtiin tämän osion kanssa, esitän tässä erittäin lyhyen johdannon tilastolliseen kielentutkimukseen.

Aluksi ihan lyhyesti laadullisista ja määrällisistä menetelmistä. Nämä ovat kaikille varmasti tuttuja asioita, mutta on hyvä kuitenkin alustaa näistä lyhyesti.

Laadulliset menetelmät ovat luonteeltaan sanallisia, luokittelevia ja painottavat tutkimuskohteen sellaisia puolia, joita ei oikein millään tavalla voi mitata tai joita on erittäin hankala mitata. Tavoitteena laadullisessa tutkimuksessa on tutkimuskohteen ymmärtäminen, ei niinkään yleistysten tekeminen. Laadullisessa tutkimuksessa tieto on usein tulkitsijasta ja kontekstista riippuvaista ja tässä mielessä jossain määrin subjektiivista. Aineisto tyypillisesti voi pohjautua vaikkapa tapaustutkimukseen.

Määrällisessä tutkimuksessa sen sijaan pyritään tekemään yleistyksiä aineiston perusteella tutkimuskohteesta ja sitä kautta ymmärtämään tutkimuskohdetta paremmin. Määrälliset menetelmät perustuvat numeeriseen aineistoon tai sellaiseen aineistoon, joka voidaan muuttaa numeeriseksi. Tämän vuoksi määrällisiä menetelmiä käytettäessä on välttämätöntä ymmärtää matematiikkaa, etenkin tilastomatematiikkaa. Tässä metodipankin osiossa käsitellään myös jonkin verran tilastomatematiikkaa, mutta vain pintapuolisesti, jotta voidaan ymmärtää esimerkiksi khiin neliön toimintaperiaatetta.

On hyvä muistaa, kun käytetään määrällisiä menetelmiä, että kaikkea ei voi mitata ja niitäkin asioita, mitä voidaan mitata, niin tyypillisesti niihin täytyy ottaa jollakin tavalla yksinkertaistettu näkökulma, jotta voidaan tarkastella niitä muuttujia tilastollisin menetelmin.

Tilastolliset menetelmät jaetaan karkeasti kahteen ryhmään yleensä. Ensinnäkin on kuvailevat menetelmät ja toiseksi tilastolliset päättelymenetelmät. Kuvailevia menetelmiä ovat mm. aineiston tunnusluvut kuten mediaani ja keskiarvo ja erilaiset visuaaliset esitystavat, joihin käytetään kaavioita tai diagrammeja esittämään aineisto jotenkin visuaalisesti, jotta siitä saa nopeasti havainnollisen käsityksen. Tilastollisen päättelyn

menetelmät, niitten avulla puolestaan tehdään päätelmiä tutkittavasta ilmiöstä aineistoon pohjautuen. Tällöin arvioidaan esimerkiksi muuttujien välistä korrelaatiota tai riippuvuutta.

Aivan lyhyesti tällaisen määrällisen tutkimuksen historiasta kielentutkimukseen sovellettuna. Voisi sanoa, että jonkinlaisia aihioita määrällisestä suuntautumisesta voidaan nähdä jo 1200-luvulta lähtien, kun alettiin kehitellä ensimmäisiä Raamatun sanojen konkordansseja. Varsinaiset tilastolliset lähestymistavat kehittyivät kuitenkin 1900-luvun puolivälin jälkeen kielentutkimuksessa. Voisi sanoa, että ensimmäisten joukossa sosiolingvistiikassa ja kokeellisessa tutkimuksessa. 1960-luvun lopulta lähtien jo sosiolingvistiikassa sovellettiin logistista regressiota William Labovin kehittelemässä Varbrul-menetelmässä, mikä on sikäli huomionarvoista, että esimerkiksi logistista regressiota on käytetty ja sovellettu vaikkapa korpustutkimuksessa selvästi myöhemmin ja vaikkapa kielitypologiassa vielä selvästi myöhemmin, vasta aivan parin viime vuosikymmenen aikana.

Noam Chomskyllä on ollut oma roolinsa myös tilastollisen lähestymistavan kehittämisessä kielentutkimukseen. Ei niinkään sillä tavalla, että hän olisi itse sitä kehittänyt tai että generatiivisessa tutkimuksessa olisi sovellettu niinkään tilastollista lähestymistapaa, vaan koska hän vastusti esimerkiksi korpustutkimusta niin kiivaasti, niin se tietenkin johti sitten siihen, että toiset innostuivat empiirisestä korpustutkimuksesta ja tilastollisen lähestymistavan soveltamisesta mm. korpuksiin.

Nykyään tilastomenetelmiä käytetään yhä enenevässä määrin myös kielentutkimuksessa ja myös sellaisilla aloilla, jotka perinteisesti kielentutkimuksessa on olleet paljolti laadullista tutkimusta kuten diskurssintutkimus, jota nykyään voi edes jossain määrin myös lähestyä määrällisin menetelmin.

Tästä metodipankin osiosta kerron lyhyesti seuraavassa. Tässä osiossa käymme läpi määrällisen tutkimuksen tutkimusprosessin askel askeleelta sovellettuna kielentutkimukseen. Eli käymme läpi tutkimuskysymyksen, tutkimushypoteesin muodostamisen, aineiston laadinnan, otantaan liittyviä asioita, sopivan menetelmän valinnan, tilastollisen testaamisen, tilastollisen merkitsevyyden tulosten tulkinnan ja lopuksi myös tutkimusraportin kirjoittamisen. Kaikki analyysi tässä osiossa tehdään käyttäen Exceliä, ja kaikki analyysit havainnollistetaan myös Excelissä. Käytännöllisiä taitoja, joita tässä osiossa opitaan, liittyy mm. aineiston ristiintaulukointiin, khiin neliö -testiin ja kuvaajien piirtämiseen. Seuraavaksi käydään sitten tarkemmin tutkimusprosessin eri vaiheisiin.