HISTORIALLINEN KORPUSLINGVISTIIKKA (TANJA SÄILY)
TRANSKRIPTIO

Hei! Tervetuloa Kielentutkimuksen metodipankkiin ja historiallisen korpuslingvistiikan pariin. Minä olen Tanja Säily ja toimin englannin kielen apulaisprofessorina Helsingin yliopistossa.

Kerron tässä lyhyesti siitä, mitä historiallinen korpuslingvistiikka on ja millaisia aiheita sillä voidaan tutkia. Kun olet katsonut tämän videon ja tehnyt siihen liittyvät tehtävät, osaat kuvailla historiallisen korpuslingvistiikan perusteita, etuja ja rajoitteita, joista on syytä olla tietoinen ennen tutkimuksen aloittamista.

Miten kieli muuttuu? Jaakko-kuninkaan aikaan 1600-luvulla englannin kielessä voitiin vielä sanoa ”The Lord giveth and the Lord taketh away”, vaikka tämä *taketh*-muoto oli jo silloin vanhahtava. Seuraavalla vuosisadalla oltiin jo siirrytty käytännössä kokonaan *-s*-muotoon, kuten tässä Hester Piozzi: ”The Fog takes away all Taste of going out”. Muun muassa tällaisten muutosten kulkua voidaan tutkia historiallisen korpuslingvistiikan avulla.

Historiallinen korpuslingvistiikka on metodiikka, jota käytetään eri kielten historian ja muutoksen tutkimiseen. Se ei siis ole oma kielitieteen haaransa, vaan historiallisen korpuslingvistiikan menetelmiä voidaan käyttää monien eri kielitieteen osa-alueitten tutkimuksessa kognitiivisesta kielitieteestä sosiolingvistiikkaan.

Aineistona historiallisessa korpuslingvistiikassa käytetään sähköisessä muodossa olevia tekstikokoelmia. Mitä tahansa elektronista tekstikokoelmaa ei kuitenkaan voida pitää korpuksena, vaan ainakin kahden kriteerin tulee täyttyä. Ensinnäkin kokoelman tulee olla riittävän laaja, jotta määrällinen tutkimus olisi mahdollista. Toiseksi kokoelman tulee olla koottu systemaattisten periaatteiden mukaisesti. Sen pitäisi sisältää sopivassa suhteessa erityyppisiä tekstejä, ja tekstien pitäisi edustaa tietyn aikakauden ja kieliyhteisön käyttämää kieltä eri konteksteissaan.

Tämä korpuksen edustavuus ja käyttökontekstien välinen tasapaino ovat korpuslingvistisessä tutkimuksessa ensiarvoisen tärkeitä, koska niitten avulla varmistetaan tuloksien luotettavuus ja yleistettävyys. Etenkin historiallisessa korpuslingvistiikassa edustavuuden ja tasapainon saavuttaminen on kuitenkin vaikeaa, ja tästä puhunkin lisää myöhemmin.

Historiallisessa korpuslingvistiikassa yhdistyvät määrällinen ja laadullinen näkökulma. Pelkät luvut eivät ole vielä mikään tulos, vaan niitä pitää pystyä tulkitsemaan. Siksi tutkijan on tunnettava käyttämänsä korpus ja sen historialliset kontekstit mahdollisimman hyvin.

Mitä historiallisen korpuslingvistiikan avulla voidaan tutkia? Tutkimusaiheet voidaan jakaa kahteen ryhmään: mikro- ja makrotasoon. Mikrotaso keskittyy yksittäisten kielen ilmiöitten vaihteluun ja muutokseen. Nämä ilmiöt voivat liittyä mihin tahansa kielijärjestelmän osa-alueeseen: fonologiaan, morfologiaan, syntaksiin, semantiikkaan, pragmatiikkaan ja niin edelleen. Esimerkkejä tutkittavista ilmiöistä ovat taivutuspäätteet, apuverbirakenteet ja sanajärjestys.

Makrotaso puolestaan pureutuu laajempiin kysymyksiin kielenmuutoksesta. Näihin kuuluvat muun muassa se, mitkä kielensisäiset ja -ulkoiset tekijät vaikuttavat muutokseen. Kielenulkoisista tekijöistä on esimerkiksi arveltu, että kielenmuutos on nopeampaa yhteisöissä, joissa ihmisten väliset siteet ovat heikkoja, kuten suurissa kaupungeissa. Hitaampaa muutos on pikkukylissä, joissa kaikki tuntevat toisensa hyvin. Voidaan myös kysyä, mitkä ihmisryhmät tyypillisesti johtavat kielenmuutosta. Aika paljon evidenssiä on löytynyt sille, että muutoksia johtavat usein naiset, mutta tästä on toki poikkeuksia. Kun itse tutkin englannin uudissanoja, havaitsin, että niitä käyttivät 1600- ja 1700-lukujen kirjeissä eniten keski-ikäiset miehet.

Käytännössä mikro- ja makrotaso monesti yhdistyvät tutkimuksessa: keskitytään tiettyyn ilmiöön ja tarkastellaan sen muutokseen vaikuttavia kielensisäisiä ja/tai kielenulkoisia tekijöitä.

Palataan esimerkkiin englannin yksikön kolmannen persoonan verbipääte *-s*:stä. Tässä on kyseessä mikrotason ilmiö, mutta siitä voidaan kysyä myös makrotason kysymyksiä, kuten että ketkä johtavat muutosta. Tästä viivadiagrammista näkyy, että naiset ovat johtaneet muutosta muuten paitsi ihan alussa, 1400-luvulla. Tätä *-s*-muodon yleistymistä ovat tutkineet Terttu Nevalainen ja Helena Raumolin-Brunberg Helsingin yliopistosta. He huomasivat, että kyseessä oli oikeastaan kaksi eri muutosta, mikä selittää 1400-luvun omituisuuden, mutta ei nyt mennä siihen. Sukupuolen lisäksi merkittäviksi *-s*-muodon käyttöä selittäviksi tekijöiksi nousivat kirjoittajien murretausta – muutos alkoi Pohjois-Englannista – sekä yhteiskunnallinen asema: nopeita omaksujia olivat etenkin säätyläisiin kuulumattomat ja toisaalta sosiaaliset nousijat. Lisätietoja tästä muutoksesta tulkintoineen löytyy Tertun ja Helenan kirjasta.

Historialliseen korpuslingvistiikkaan liittyy paljon hyviä puolia. Ensinnäkin sen avulla voidaan saada yleiskuva tutkittavan ilmiön vaihtelusta ja muutoksesta. Tämä muodostetaan tyypillisesti tarkastelemalla ilmiön esiintymistiheyttä eli frekvenssiä ajassa. Puhun myöhemmin lisää

frekvenssin laskemisesta. Sen avulla päästään käsiksi kielenkäytön tendensseihin: monesti on niin, että joku ilmiö ei esiinny kategorisena joko–tai-jakona, vaan pikemminkin tendenssinä, joka lopulta voi aiheuttaa kielioppijärjestelmän muuttumisen. Esimerkiksi johtimien morfologinen produktiivisuus eli se, missä määrin niitä käytetään sanojen muodostamiseen, ei ole mikään joko–tai-kysymys, vaan voi olla enemmän tai vähemmän produktiivisia suffikseja. Produktiivisuuden muutos ajassa on vähittäistä mutta voi lopulta johtaa uuden suffiksin vakiintumiseen tai vanhan suffiksin käytön loppumiseen.

Historiallisen korpuslingvistiikan toinen hyvä puoli on sen aineistolähtöisyys. Me emme voi matkustaa satoja vuosia ajassa taaksepäin kysymään ihmisiltä, miten he jonkin asian ilmaisevat, vaan meidän on pakko käyttää heiltä säilyneitä tekstejä. Niistä voidaan saada selville, miten kieltä oikeasti käytettiin, sen sijaan että luotettaisiin kielenpuhujien intuitioon. Toki aineistoilla on myös rajoitteensa, joista kerron lisää kohta.

Menetelmän kolmas hyvä puoli on sen kieliriippumattomuus. Jos kielestä on saatavilla riittävästi aineistoa, niin sitä voidaan periaatteessa käsitellä samoilla keinoilla kuin muittenkin kielten aineistoja. Käytännössä kaikki työkalut eivät kuitenkaan tue kaikkia eri kirjoitusjärjestelmiä, ja joissakin kielissä jopa sanan määritteleminen voi olla vaikeampaa kuin toisissa, mikä vaikeuttaa sanamääriin perustuvaa frekvenssien laskemista. Yleensä ottaen menetelmä kuitenkin toimii hyvin, ja yleisiä kielen ilmiöitä kuten vaikka apuverbejä voidaan tutkia melko pienelläkin aineistomäärällä.

Historiallisen korpuslingvistiikan aineistolähtöisyys voidaan nähdä sekä vahvuutena että heikkoutena. Muilla voi olla big dataa, mutta historiallisilla lingvisteillä onkin bad dataa, josta on kirjoittanut muun muassa tunnettu sosiolingvisti William Labov.

Historiallinen korpusaineisto on enimmäkseen kirjallista, koska puheen tallentamiseen tarvittava teknologia on suhteellisen uutta, eikä kielitieteelliseen tutkimukseen soveltuvaa historiallista puheaineistoa ole vielä ehtinyt kertyä. Varhaisemmilta vuosisadoilta on kyllä saatavilla kirjurien muistiin kirjoittamia oikeudenkäyntejä ja muita vastaavia. Kuten aiemmin totesin, meillä ei ole myöskään pääsyä puhujien intuitioihin. Jonkin verran kuitenkin löytyy niin sanottua metalingvististä diskurssia, jossa ihmiset kirjoittavat mielipiteitään vaikkapa eri alueilla puhutuista murteista viitaten tiettyihin kielenpiirteisiin.

Aineisto tuppaa myös olemaan varsin niukkaa: monien kielten kirjoitettu historia on lyhyt, tekstejä ei ole välttämättä alun perinkään tuotettu kovin paljon, ja niitä on säilynyt vielä vähemmän. Niistäkin teksteistä, jotka ovat meidän päiviimme asti säilyneet, usein vain pieni osa

on digitoitu korpuksiin. Myös taustatieto tekstien historiallisista konteksteista on yleensä vajavaista, ja niitten kirjoittajat jäävät tyypillisesti hämärän peittoon. Historiantutkimus voi kuitenkin auttaa meitä ymmärtämään näitä kontekseja. Itse sosiolingvistinä olen tutustunut uuden ajan Englannin sosiaalihistoriaan, jotta ymmärtäisin, miten tuon ajan yhteiskunnallinen hierarkia on saattanut vaikuttaa yksilöitten kielenkäyttöön.

Yksi suurimmista ongelmista on aineiston epätasaisuus. Aineistoa on tuotettu, säilynyt ja digitoitu epätasaisesti eri yhteiskuntaryhmistä, tekstilajeista, murteista ja niin edelleen. Näitten väliset suhteet vaihtelevat aikakausittain, mikä tietysti vaikeuttaa ajallista vertailua. Yhteiskuntaryhmistä hyväosaiset miehet ovat perinteisesti olleet kirjoitustaitoisimpia, ja heidän kirjoituksiaan on pidetty tärkeinä säilyttää ja julkaista. Painotekstejä taas on säilynyt paremmin kuin vaikkapa yksityiskirjeitä, ja niitten digitointikin on helpompaa. Lisäksi eri aikakausina on ollut olemassa erilaisia tekstilajeja, esimerkiksi sanomalehdet kehittyivät oikeastaan vasta 1700-luvun aikana. Murteista kauhuesimerkiksi voidaan ottaa 1000-luvun molemmin puolin puhutut muinais- ja keskienglanti: suurin osa muinaisenglannin ajalta säilyneistä teksteistä on kirjoitettu Etelä-Englannissa puhutulla Wessexin murteella, kun taas keskienglantiin siirryttäessä tekstit tulevatkin suurimmaksi osaksi pohjoisemmalta Midlandsin alueelta. Tutkipa siinä sitten kielenmuutosta, kun vertaat samalla vähän niin kuin 1800-luvun Turun murretta 2000-luvun Kuopion murteeseen!

Vaikka näitä rajoitteita onkin paljon, niitä ei kannata pelästyä – ne eivät mitenkään estä tutkimuksen tekemistä, mutta ne pitää ottaa huomioon kaikissa tutkimuksen vaiheissa aineiston valinnasta tulosten tulkintaan.

Niissä teksteissä, jotka on kirjoitettu ennen kielen standardisaatiota, on hyvin merkittävä käytännön ongelma: sanojen kirjoitusasujen vaihtelu. Se hankaloittaa kielen ilmiöitten hakemista tekstistä, paitsi jos tutkimuskohteena on juuri ortografia. Otan vaihtelusta esimerkiksi William Shakespearen nimen, mutta sama pätee kyllä ihan kaikkiin sanoihin.

Niin kuin näistä kuvista näkyy, näytelmäkirjailija itse kirjoitti omaa nimeään ainakin kuudella eri tavalla. 1600- ja 1700-lukujen teksteistä löytyy sukunimelle reilu tusina muutakin varianttia. Ja Shakespearen nimi vakiintui nykyiseen kirjoitusasuunsa vasta 1900-luvulla.

Nyt päästän sinut tekemään tähän osioon liittyviä tehtäviä. Nähdään seuraavassa osiossa!