VIDEO 2: HISTORIALLISET KORPUKSET (TANJA SÄILY)
TRANSKRIPTIO

Hei! Tervetuloa taas historiallisen korpuslingvistiikan pariin. Minä olen Tanja Säily ja tämä video kertoo siitä, millaisia historiallisia korpuksia on olemassa ja miten niihin pääsee käsiksi.

Tässä heti alkuun yksi esimerkki historiallisesta korpuksesta. Corpus of Early English Correspondence on englannin kirjekorpus, jolla voidaan tutkia sosiolingvistisiä kysymyksiä, kuten sitä, ketkä johtavat muutosta. Kuvassa korpus on sovitettu kokeelliseen Khepri-käyttöliittymään, jossa näkyy tekstin lisäksi reaaliaikaisesti tuotettuja kuvia muutoksesta. Vasemmalla on tehty haku, tässä tapauksessa has-verbistä eri kirjoitusasuineen, ja keskellä on niin sanottu ”key word in context”- eli KWIC-konkordanssilistaus, jossa näkyy hakusanan osumat ja niitten ympärillä kontekstia. Osumaa klikkaamalla saa näkyviin pitemmän kontekstin. Tällainen konkordanssitoiminnallisuus löytyy lähes kaikista korpustyökaluista.

Korpus on siis elektroninen tekstikokoelma, joka on riittävän iso määrälliseen tutkimukseen ja joka on koottu systemaattisten periaatteitten mukaisesti. Se pyrkii edustamaan jonkin kieliyhteisön tietyissä konteksteissa käyttämää kieltä, ja tekstit pyritään valitsemaan tasapainoisessa suhteessa näistä konteksteista. Sosiolingvistisessä tutkimuksessa korpuksen pitää edustaa mahdollisimman tasapainoisesti myös kieliyhteisön sisäisten sosiaaliryhmien, kuten naisten ja miesten, kielenkäyttöä.

Kielen käyttökontekstit määritellään usein genreinä eli tekstilajeina. Jos korpuksen on tarkoitus edustaa yhteisön kielenkäyttöä mahdollisimman laajasti, siihen valitaan tekstejä sopivassa suhteessa useista genreistä ja sitä kutsutaan yleiskorpukseksi. Vaihtoehtoisesti voidaan keskittyä vain yhteen genreen, esimerkiksi jos kiinnostuksen kohteena on nimenomaan tuon tekstilajin kehitys. Yhden genren korpuksellakin voidaan tarvittaessa tutkia koko kielen muutosta, mutta silloin pitää pohtia hyvin tarkkaan, miten yleispäteviä tulokset oikeastaan ovat.

Historialliset korpukset voivat olla joko synkronisia tai diakronisia. Synkroninen korpus edustaa vain yhden ajankohdan kieltä, ja sillä voidaan tutkia vaikkapa kielenkäytön genrevaihtelua tuona ajankohtana. Diakroniseen korpukseen on koottu tekstejä mahdollisimman tasapainoisesti useilta eri aikakausilta, ja sillä voidaan tutkia vaihtelun lisäksi kielen ajallista muutosta.

Otetaan esimerkkejä näistä erityyppisistä korpuksista.

Brownin yliopistossa 1960-luvulla koottu Brown Corpus edustaa aikansa kirjoitettua amerikanenglantia. Se on siis synkroninen, ja nykyään sitä voidaan pitää historiallisena, kun 60-luvusta on jo kulunut sen verran aikaa. Korpukseen on otettu useista eri genreistä yhteensä 500 tekstinäytettä, jotka ovat noin 2000 sanan mittaisia. Tekstien yhteenlaskettu sanamäärä on siten miljoona. Niistä neljännes on fiktiota ja kolme neljäsosaa ei-fiktiota, kuten akateemista ja sanomalehtitekstiä. Tämä tasapaino ei välttämättä ole sama kuin reaalimaailmassa, mutta ideana on ollut, että aineistoa olisi joka kategoriasta riittävästi määrällistä vertailua varten.

Samoilla periaatteilla on sittemmin koottu muitakin korpuksia, niin että nykyään korpusperhe kattaa sekä britti- että amerikanenglantia kolmelta eri vuosikymmeneltä. Niinpä sillä voidaan tehdä myös diakronista tutkimusta. Yhtenä ongelmana on kuitenkin genrejen ajallinen vertailukelpoisuus. Huomaatko taulukossa luetelluissa tekstilajeissa mitään erikoista?

Esimerkiksi fiktion alta löytyvä seikkailu- ja länkkärikategoria on varmaan ollut 60-luvulla hyvinkin kurantti, mutta liekö näin enää 90-luvulla? Lisäksi 90-luvulla oli ilmaantunut uusia nettigenrejä, joista 60-luvulla ei ollut vielä tietoakaan, joten korpus edustaa siinäkin mielessä huonosti 90-luvun genrejä. Tällaiset ongelmat tietysti kertaantuvat pitempiä ajanjaksoja edustamaan pyrkivissä korpuksissa.

Helsinki Corpus of English Texts on täällä Helsingissä koottu englannin diakroninen yleiskorpus. Se kattaa peräti 1000 vuotta kirjoitettua englantia useista eri genreistä, 700-luvulta 1700-luvulle. Helsinki Corpuksen suhteen onkin tehty eri ratkaisu kuin Brownin: jos jollakin aikakaudella on ilmaantunut uusi tärkeä tekstilaji, se on otettu mukaan korpukseen, vaikka tämä muuttaakin korpuksen koostumusta ajassa. Tässä on yksinkertaistettu kuva korpuksen tekstikategorioista keski- ja varhaisuusenglannin kausilla (muinaisenglanti on tästä jätetty pois). Esimerkiksi lakitekstejä kirjoitettiin normannivalloituksen jälkeisinä vuosisatoina lähinnä latinaksi ja ranskaksi, ja englanninkielistä materiaalia alkaa löytyä vasta myöhäiskeskienglannista alkaen.

Helsinki Corpus oli yksi varhaisimmista historiallisista korpuksista, ja professori Matti Rissanen tiimeineen teki sen kanssa aivan uraauurtavaa työtä. Korpusta on sittemmin käytetty ympäri maailmaa, ja sen pohjalta on kirjoitettu satoja tieteellisiä artikkeleita. Monet Matin tiimin jäsenistä ovat jatkaneet historiallisten korpusten kokoamista. Yksi näistä korpuksista on jo aiemmin mainittu Corpus of Early English Correspondence.

Corpus of Early English Correspondence eli CEEC on diakroninen yhden genren korpus, joka sisältää henkilökohtaisia kirjeitä neljänsadan vuoden ajalta. Kirjeet on valinnut ja digitoinut

painetuista kirje-editioista professori Terttu Nevalaisen johtama tiimi täällä Helsingin yliopistossa. Korpus on jonkin verran isompi kuin Helsinki Corpus, ja sen vahvuus on sosiaalisessa edustavuudessa: koska kirjeitä kirjoittivat kaikki kirjoitustaitoiset, niitä pystyttiin kokoamaan korpukseen kaikilta yhteiskuntaryhmiltä palvelusväestä kuninkaallisiin. Kirjeitä on kuitenkin ollut saatavilla epätasaisesti, joten sosiaaliryhmien tasapaino vaihtelee ajassa. Eniten kirjeitä löytyy hyväosaisilta miehiltä, mutta naisiakin on korpuksen kirjoittajista noin neljännes. Kirjoittajien taustatiedot on koottu erilliseen tietokantaan, ja tätä ns. metadataa voidaan käyttää sosiolingvistisessä analyysissa. Metadata on tietoa datasta, eli tässä tapauksessa korpukseen liittyvää taustatietoa.

Edelliset esimerkit olivat kaikki niin sanottua rikasta dataa: korpukset ovat suhteellisen pieniä ja huolellisesti koottuja, ja niihin on liitetty paljon metadataa tekstilajeista tai sosiaaliryhmistä. Esimerkiksi CEEC-korpuksen viisi miljoonaa sanaa vastaavat noin viittäkymmentä romaania, eli se olisi vielä periaatteessa ihmisen luettavissa kokonaan. Nykyään historiallisista tekstiaineistoista alkaa olla saatavilla myös niin sanottua isoa dataa, jota yksittäinen ihminen ei pysty tutkimusta varten lukemaan kokonaan.

On ensinnäkin koottu satojen miljoonien sanojen megakorpuksia. Ne pyrkivät kyllä edustavuuteen ja tasapainoon tekstilajien suhteen, mutta koska aineistoa on niin paljon, että sen tarkka käsittely on ihmisvoimin mahdotonta, tekstit ja metadata on koottu koneellisesti ja niissä voi olla virheitä. Näitten korpusten etuna on se, että niitten avulla voidaan tutkia myös harvinaisempia kielen ilmiöitä, ja muutoksia voidaan tutkia ajallisesti tarkemmin, joskus jopa vuositasolla. Esimerkiksi Brigham Young -yliopiston professori Mark Davies on koonnut tällaiset korpukset espanjasta ja englannista, sekä vähän pienemmän korpuksen portugalista.

Lisäksi on saatavilla digitoituja tekstitietokantoja, jotka eivät ole lingvistien kokoamia, joten niitten kokoamisessa ei ole kiinnitetty samanlaista huomiota kielelliseen edustavuuteen ja tasapainoon. Esimerkiksi Early English Books Online eli EEBO sisältää valtavan määrän uuden ajan englanninkielisiä painotekstejä. Tekstit on valikoitu kaupallisen ProQuest-yhtiön ja yli 150 kirjaston yhteistyönä erilaisten bibliografioitten perusteella. EEBOn nykyisen version tekstit on digitoitu sen verran huolella, että niissä ei ole paljon virheitä, mutta metadata on peräisin niin monesta eri kirjastokatalogista, että sen muoto ja laatu vaihtelevat tosi paljon.

Kansalliskirjaston lehtikokoelmassa onkin sitten jo yli viisi miljardia sanaa, tai enemmänkin jos lasketaan kaikki kielet ja nekin osat joita ei ole vielä saatavilla korpuskäyttöliittymässä. Tämä viiden miljardin sanan osuus edustaa suomenkielistä sanomalehtitekstiä varsin kattavasti, joskin tasapainon suhteen on sanottava, että myöhemmiltä ajoilta on tietysti valtavan paljon

enemmän lehtiä kuin varhaisemmilta. Lisäksi automaattinen digitointi on jättänyt tekstiin paljon virheitä, jotka vaikeuttavat analyysia.

Tästä kuvasta näkyy aineistojen kokoerot. Viiden miljoonan sanan kirjekorpus näkyy tässä vain hiuskarvan paksuisena, ja satojen miljoonien sanojen megakorpus ja kirjatietokantakin kalpenevat miljardien sanojen lehtitietokannan rinnalla.

Niin rikkaalla kuin isollakin datalla on paikkansa kielentutkimuksessa. Lisätietoa niistä löytyy muun muassa Turo Hiltusen ja kumppanien kirjoittamasta artikkelista.

Historiallisia korpuksia löytyy usein samoista paikoista kuin muitakin korpuksia. Suomen Kielipankkiin on sijoitettu paljon kotimaisten ja suomensukuisten kielten korpuksia, mutta sen kielivalikoima kattaa myös muun muassa venäjän ja englannin. Lisää korpuksia voi löytää erilaisten metatietopalvelujen kautta. Näitä ovat esimerkiksi eurooppalaiset META-SHARE ja Language Resource Inventory, amerikkalainen Linguistic Data Consortium sekä täällä Helsingissä ylläpidettävä Corpus Resource Database eli CoRD, joka on keskittynyt englanninkielisiin korpuksiin. Aina kannattaa tarkistaa myös oman yliopiston aineistokokoelmat, joista saattaa löytyä unohdettuja helmiä tai ilmainen pääsy johonkin maksulliseen aineistoon. Jos sopivaa korpusta ei ole olemassa, voi sellaisen koota myös itse, mutta tämä voi olla hyvin työläs prosessi.

Sitten kun korpus on löytynyt, tarvitaan joku ohjelma, jolla siitä voi tehdä hakuja. Jotkut korpukset ovat saatavilla web-käyttöliittymässä. Se voi olla räätälöity kyseiselle korpukselle: esimerkiksi Mark Davies on kehittänyt oman käyttöliittymän kokoamilleen englannin, espanjan ja portugalin korpuksille. Käyttöliittymä voi olla myös yleisempi, niin kuin pohjoismaisten kielipankkien käyttämä Korp tai maailmalla yleinen CQPweb. Lancasterin yliopiston CQPweb-palvelimen kautta pääsee käsiksi moniin englannin ja muittenkin kielten korpuksiin, kunhan sinne rekisteröityy yliopiston sähköpostiosoitteella.

Omalle koneelle ladattua korpusta voidaan käyttää myös erillisellä konkordanssiohjelmalla. Näitä ovat esimerkiksi ilmainen AntConc ja maksullinen WordSmith Tools. Esittelen AntConcin käyttöä tarkemmin seuraavassa videossa.

Kolmas vaihtoehto korpuksen käsittelyyn on ohjelmointikieli, kuten Python tai R-ympäristö. Tämä vaatii tietysti ohjelmointitaitoa, mutta etuna on se, että tutkija ei ole sidottu korpustyökalusta valmiiksi löytyviin toimintoihin, ja toisaalta tutkimus on periaatteessa helposti toistettavissa, kun kuka tahansa voi ajaa tutkijan käyttämän skriptin.

Näillä eväillä päästäänkin jo tekemään historiallista korpustutkimusta. Nähdään seuraavassa videossa!