

VIDEO 3a. AINEISTON HAKEMINEN ANTCONCILLA (TANJA SÄILY)

TRANSKRIPTIO

Tervetuloa tekemään historiallista korpuslingvistiikkaa! Minä olen Tanja Säily ja näillä videoilla käydään läpi tutkimusprosessi aineiston hakemisesta ja koodaamisesta analyysiin ja tulkintaan. Konkordanssiohjelmana käytetään AntConcia ja taulukkolaskentaohjelmana Exceliä. Otetaan esimerkiksi jo aiemmin mainittu englannin 3. persoonan verbipääte *-s/-th*. Terttu Nevalainen ja Helena Raumolin-Brunberg havaitsivat, että CEEC-kirjekorpuksessa *-s*-pääte yleistyi naisten johtamana varhaisuusenglannin kaudella. Tutkitaanpa, miten muutos etenee samalla kaudella Helsinki Corpuksessa, joka on monen genren yleiskorpus. Hakemisen helpottamiseksi keskitytään yleisimpään verbiin, *has/hath*.

Olen ladannut koneelleni AntConc-ohjelman sekä Helsinki Corpuksen varhaisuusenglannin osuuden, joka on saatavilla metodipankin kautta. Avataan korpustiedostot AntConcissa valitsemalla File > Open File(s), tai jos korpustiedostot ovat kansiossa, jossa ei ole mitään muuta, voidaan käyttää myös Open Directory -komentoa. Valitsen nyt Open Dir ja navigoin kansioon, jossa korpus koneellani on ja painan Choose. AntConc näyttää korpustiedostot vasemmalla olevassa paneelissa. Niitä on näköjään 81 kappaletta.

Katsotaan ensin uutta *has*-varianttia. Nyt voidaan tehdä heti haku: kirjoitetaan *has* ja painetaan Start, jolloin AntConc näyttää hakutulokset tutussa konkordanssinäkymässä. Ongelmana on kuitenkin se, että varhaisuusenglannin oikeinkirjoitus ei ollut vielä vakiintunut, joten tämä haku ei löydä kaikkia esiintymiä. Kirjoitusasujen vaihtelu voi olla jopa niin merkittävä ongelma, että tutkimuksen kvantitatiivisiin tuloksiin ei voi luottaa, ellei sitä ole otettu asianmukaisesti huomioon jo tässä hakuvaiheessa.

Sanojen eri kirjoitusasujen löytämiseen on kaksi ratkaisua: Ensinnäkin voidaan tehdä AntConcin sisällä lista kaikista korpuksen sanoista ja selailta sitä. Toiseksi niitä voidaan etsiä historiallisista sanakirjoista, englannin kielen tapauksessa Oxford English Dictionarysta eli OED:sta. Katsotaan ensin AntConcin sanalistatoimintoa. Se löytyy Word List -välilehdeltä. Tässä ei tarvitse muuta kuin painaa Start. AntConc järjestää sanalistan oletuksena yleisyyden mukaan, mutta meitä kiinnostaa saman sanan eri kirjoitusasut, joten järjestetään sanat aakkosjärjestykseen: valitaan Sort by -tekstin alta Sort by Word ja painetaan Sort. Nyt sanat ovat aakkosjärjestyksessä ja voidaan hypätä *has*-

sanan tienoille painamalla Search Only. Sieltä löytyy muoto "hase", josta veikkaan, että se voisi edustaa *has*-sanaa. Klikataan sitä, niin nähdään onko näin. Hyvältä näyttää!

Ruvetaan keräämään näitä kirjoitusasuja erilliseen tekstitiedostoon, kukin muoto omalle rivilleen. Voit tehdä uuden tekstitiedoston esimerkiksi Notepad-ohjelmassa ja tallentaa sen haluamaasi paikkaan. Itse käytän TextMate-nimistä ohjelmaa Macille. Muotoja oli siis tähän mennessä *has* ja *hase*. Mennään takaisin sanalistaan ja katsellaan, löytyisikö niitä lisää. En ainakaan heti huomaa mitään lupaavaa. Siirrytään sitten OED:hen.

Haetaan täällä *has*-sanaa, joka löytyy *have*-verbin alta. Eri kirjoitusasut löytyvät Forms-kohdasta klikkaamalla Show More. Skrollataan yksikön kolmanteen persoonaan ja *has*-varianttiin. Täällä onkin paljon tavaraa, mutta keskitytään nyt varhaisuusenglannin muotoihin, 1500-luvulta 1700-luvulle. Täältä löytyy ainakin *hace* c:llä, otetaan nyt myös tuo yorkshirelaismuoto *hays*, sitten onkin jo normaali *has*, *ha's* heittomerkillä ja pohjoismurteitten muoto *hes*. Täällä on vielä jotain skottimuotoja, mutta jätetään ne nyt pois ja keskitytään englanninenglantiin.

Näistä tulikin jo kiva lista. Tästä voi vielä etsiä säännönmukaisuuksia, joitten perusteella voi keksiä lisää mahdollisia muotoja. Esimerkiksi s-elementin jälkeen näyttäisi tulevan usein e-kirjain, joten voidaan arvata, että nuo kolme viimeistä muotoa voitaisiin ehkä kirjoittaa joskus e:llä. Lisäksi s voidaan joskus kirjoittaa c:nä, mutta eiköhän tämä lista riitä meidän tarpeisiin. Jos olisin tekemässä ihan oikeaa tutkimusta, niin lisäisin varmuuden vuoksi myös c:lliset ja heittomerkilliset muodot ja selailisin AntConcin sanalistaan vähän lisää näitten sanojen tienoilta. Kirjoitusasujen vaihtelua on paljon etenkin käsin kirjoitetuissa yksityisteksteissä, kuten kirjeissä, joita olen itse tutkinut. Muista tallentaa lista! Hakutermien tallentaminen on siksikin tärkeää, että ne pitää aina raportoida opinnäytetyössä tai tieteellisessä artikkelissa. Usein tutkimustyössä myös kestää niin pitkään, että tutkija ehtisi itsekin unohtaa, mitä on tehnyt, jos ei olisi dokumentoinut sitä.

Tehdään sitten haku listan sanoilla. Painetaan AntConcin konkordanssivälilehdellä Advanced, valitaan "Use search terms from list below", ladataan siihen tuo äsken tehty tiedosto ja painetaan Apply. Sitten painetaan Start. Hakutuloksia tulee aika liuta. Järjestetään ne hakusanan mukaan eli kohdasta 0, sitten ensimmäisen oikealla puolella olevan sanan mukaan eli 1R ja sitten toisen eli 2R. Painetaan Sort. Nyt ne alkavat olla mukavassa järjestyksessä, ensin paljon *has*-muotoa, joskin tuolla välissä on yksi *has*t, joka on toisen persoonan muoto, ja muutama e:llinen ja *hes*-muoto. Koko tekstin

saa näkyviin klikkaamalla haluamaansa esiintymää. *Hes*-muodot näyttävät itse asiassa liittyvän *his*-persoonapronomiiniin, mutta se ei haittaa, niistä päästään eroon seuraavassa vaiheessa. Takaisin konkordanssilistaukseen pääsee painamalla konkordanssivälilehteä. Otetaan vielä vähän enemmän kontekstia näkyviin eli kasvatetaan hakuikkunan kokoa 50:stä sataan ja painetaan uudestaan Start. Tämä voi auttaa analyysissa. Nyt voidaan tallentaa nämä hakutulokset ja jatkaa niiden käsittelyä Excelissä. Valitaan File > Save Output tai Ctrl+S ja tallennetaan tiedosto halutulla nimellä haluttuun paikkaan. Kaikki tutkimukseen liittyvät tiedostot kannattaa pitää samassa paikassa.