VIDEO 3b. AINEISTON KÄSITTELY EXCELISSÄ (TANJA SÄILY)
TRANSKRIPTIO

Aloitetaan hakutuloksien käsittely Excelissä. Käytän tässä vuoden 2020 englanninkielistä Exceliä Macille. Ohjelman eri versioissa toiminnot voivat olla vähän eri paikoissa, mutta ne löytyvät kyllä helposti Excelin Help-valikosta tai nettihaulla. Avataan tiedosto valitsemalla File > Open, navigoimalla tallennuspaikkaan ja tuplaklikkaamalla tiedostoa. Excel osaa muotoilla tiedoston suoraan oikein, eli valitaan tässä importointi-ikkunassa vaan Finish. Nyt tiedosto on saatu Excel-taulukkoon. Sitten voitaisiin oikeastaan heti tallentaa tämä Excel-muodossa. Valitaan File > Save As, vaihdetaan tiedostomuodoksi Excel ja tallennetaan haluttuun paikkaan.

Tämä taulukko ei ole vielä kovin luettava. Levitetään sarakkeet tekstin levyisiksi valitsemalla kaikki ja tuplaklikkaamalla jotakin sarakeväliä. Sitten tehdään sarakkeille otsikot. Lisätään alkuun niille rivi: valitaan ensimmäinen rivi ja mennään Insert > Rows. Nimetään tärkeimmät sarakkeet vaikka näin: ”Vasen konteksti”, ”Oikea konteksti” ja ”Teksti”, jossa on sen tiedoston nimi, josta esiintymä on peräisin. Lihavoidaan vielä otsikkorivi valitsemalla se ja painamalla lihavointipainiketta ja kiinnitetään se niin, että se näkyy skrollatessa aina, valitsemalla sen alapuolinen rivi ja Window-valikosta Freeze Panes. Nyt taulukkoa on helpompi käsitellä.

Koska tarkoitus on tutkia *has*-muodon kehitystä ajassa, lisätään sarakkeet ”Variantti” ja ”Aikakausi”. Ensimmäiseen lisätään *has*-sana silloin kun esiintymä sitä edustaa. Jos esiintymä on roskaa, solu jätetään tyhjäksi. Koska AntConcissa jo katsottiin, että suurin osa hakutuloksista näyttäisi olevan ihan oikeita, esitäytetään koko sarake *has*-sanalla: kirjoitetaan ensimmäiseen soluun ”has” ja tuplaklikataan solun oikeaa alakulmaa, jolloin koko sarake täyttyy sillä. Nyt voidaan skrollata alaspäin ja käydä esiintymät läpi. Jos olisin tekemässä oikeaa tutkimusta, tarkistaisin joka esiintymän, koska etenkin näin pienessä korpuksessa jokainen esiintymä voi vaikuttaa tuloksiin. Tätä esimerkkiä varten riittää kuitenkin silmäily. Muuten näyttää hyvältä, mutta tällä rivillä kyseessä onkin toisen persoonan *hast*-muoto, jossa *t*-kirjain on tuollaisten sulkeitten sisällä. Nuo sulkeet tarkoittavat Helsinki Corpuksessa editoriaalista lisäystä. Poistetaan siis *has*-sana tältä riviltä. Täällä lopussa on *hase*-esiintymät, joista jo katsottiin, että ne edustavat *has*-sanaa, eli ne ovat ok. Lisäksi on kaksi *hes*-esiintymää, jotka ovat oikeasti *his*-possessiivipronomini: *his leftenant* ja *his sargeant*. Poistetaan *has*-sana näitten kohdalta.

Jos näitä *has*-esiintymiä haluttaisiin luokitella tarkemmin, niin voitaisiin lisätä uusi sarake, johon voitaisiin koodata esimerkiksi se, toimiiko *has* esiintymässä apuverbinä vai leksikaalisena verbinä. Tässä esimerkissä riittää tarkastella *has*-muotoa kokonaisuutena.

Nyt esiintymiin pitäisi lisätä aikakausi, jolloin teksti on kirjoitettu. Sen voi päätellä Helsinki Corpuksen tapauksessa tiedostonimestä: ykkönen tarkoittaa varhaisuusenglannin osion ensimmäistä kautta, kakkonen toista ja kolmonen kolmatta. Nämä kaudet on dokumentoitu korpuksen manuaalissa. Aikakaudet pystyy lisäämään riveille helposti käyttämällä Excelin suodatintoimintoa. Valitaan taas otsikkorivi ja lisätään sille suodatin menemällä Sort & Filter > Filter. Painetaan Teksti-otsikon vieressä olevaa nuolta, valitaan Filter-otsikon alta Contains ja kirjoitetaan sen viereen 1. Nyt näkyy vain ne rivit, joitten teksti on ykköskaudelta. Voidaan siis kirjoittaa Aikakausi-sarakkeeseen ykköskauden vuodet, 1500–1570. Otetaan kiinni solun oikeasta alakulmasta ja raahataan alaspäin, niin vuodet täyttyvät kaikille riveille. Sitten vaihdetaan filtteriin numero 2 ja tehdään sama juttu. Tässä vuodet ovat 1570–1640. Ja vielä kolmonen, 1640–1710. Nyt voidaan poistaa suodatin painamalla sitä ja valitsemalla Clear Filter.

Aineisto on nyt valmis analysoitavaksi. Helpoiten laskelmia voi tehdä Excelin Pivot-taulukko-toiminnon avulla. Valitaan Data > Summarise with Pivot Table ja painetaan OK. Rakennetaan sellainen taulukko, jossa riveillä on aikakaudet, eli raahataan Aikakausi-kenttä Rows-kohtaan. Sarakkeisiin halutaan variantti eli *has*-verbi. Kun meitä ei tyhjäksi jätetyt kiinnosta, ne voidaan suodattaa tuolta Column Labels -kohdasta pois. Ja laskea halutaan variantin esiintymien lukumäärää.

Taulukosta näkyy jo, että *has*-muodon esiintymistiheys eli frekvenssi kasvaa selvästi ajassa. Helsinki Corpuksessa on kuitenkin eri aikakausilta eri määrä dataa, joten näitä raakafrekvenssejä ei voi käyttää sellaisinaan, vaan ne pitää normalisoida. Lasketaan seuraavaksi normalisoidut frekvenssit.