VIDEO 3c. NORMALISOIDUN FREKVENSSIN LASKEMINEN JA VISUALISOINTI (TANJA SÄILY) TRANSKRIPTIO

Normalisoidun frekvenssin laskemiseen tarvitaan tutkittavan ilmiön esiintymien lukumäärän lisäksi tieto siitä, montako juoksevaa sanaa korpuksen kullakin aikakaudella on. Metodipankista löytyy Excel-taulukko, jossa on Helsinki Corpuksen varhaisuusenglannin osion sanamäärät ja tekstimäärät aikakausittain. Ne on dokumentoitu myös korpuksen manuaalissa, joka löytyy netistä. Jos käytät jotakin muuta korpusta, josta tietoa ei ole saatavilla, voit käyttää AntConcin sanalistatoimintoa sanamäärien laskemiseen. Silloin pitää kuitenkin ottaa huomioon se, että korpustiedostoissa on usein tekstien lisäksi metadataa, jonka AntConc laskee sanamääriin mukaan. Esimerkiksi Helsinki Corpuksen varhaisuusenglannin osion kokonaissanamäärä on 551 000 sanaa, mutta AntConcin ilmoittama sanamäärä, joka löytyy sanalistatoiminnon Word Tokens -kohdasta, on parikymmentätuhatta sanaa isompi. AntConcin asetuksia voi säätää siten, että se jättää pois tietynlaisten tägien sisältä löytyvän metadatan. Tämä toiminto löytyy valitsemalla Settings > Global Settings > Tags ja Hide tags.

Nyt sanamäärä on lähempänä oikeaa, muttei kuitenkaan vielä ihan oikein. Näin voi tapahtua, jos tägejä on monenlaisia tai jos metadataa ei ole merkitty koneluettavilla symboleilla. Tällöin käyttäjä joutuu itse arvioimaan metadatan sanamäärän ja vähentämään sen kokonaissanamäärästä.

Avataan tuo äsken mainittu Excel-taulukko ja kopioidaan sieltä aikakaudet ja sanamäärät meidän has-tiedostoon. Kopioidaan sitten Pivot-taulukosta *has*-esiintymien lukumäärä tähän samaan taulukkoon. Valitaan täältä Link Cells, niin solut päivittyvät automaattisesti jos aineiston koodausta toisella välilehdellä muutetaan myöhemmin. Normalisoinnissa lasketaan ensin, mikä osuus aikakauden juoksevista sanoista edustaa tutkittavaa ilmiötä, eli jaetaan *has*-esiintymien lukumäärä aikakauden kokonaissanamäärällä. Sitten lasketaan, montako esiintymää olisi, jos korpuksessa olisi vaikkapa satatuhatta sanaa, kertomalla tulos sadallatuhannella. Näin voidaan suoraan vertailla, mikä on *has*-sanan esiintyvyys sataatuhatta sanaa kohti kullakin aikakaudella. Normalisointikannan ei ole pakko olla aina satatuhatta, vaan korpuksen koosta ja ilmiön yleisyydestä riippuen voidaan käyttää vaikka kymmentätuhatta tai miljoonaa sanaa.

Näitten laskutoimituksien tekeminen on Excelissä helppoa. Ensin valitaan taulukon solu, johon laskelma halutaan tehdä, ja kirjoitetaan siihen yhtäsuuruusmerkki, joka kertoo Excelille, että nyt halutaan laskea jotain. Yhtäsuuruusmerkin jälkeen valitaan se solu, josta löytyy *has*-sanan esiintymien lukumäärä tällä aikakaudella. Excel lisää automaattisesti viitteen soluun. Sitten halutaan tehdä jakolasku, eli lisätään kauttamerkki, ja valitaan aikakauden kokonaissanamäärä, johon Excel lisää taas viitteen. Viimeiseksi kerrotaan tämä sadallatuhannella, eli lisätään tähti ja kirjoitetaan

satatuhatta numeroina. Painetaan enteriä ja saadaan tulos. *Has*-sanaa esiintyy 1500-luvun alussa siis vain noin viisi kertaa sataatuhatta sanaa kohti.

Tehdään sama muillekin aikakausille. Tässä voidaan käyttää Excelin täyttötoimintoa: otetaan kiinni solun oikeasta alakulmasta ja vedetään alaspäin, niin Excel toistaa laskutoimituksen kunkin rivin tiedoilla. Tulos onkin hyvin mielenkiintoinen: *has*-sanan esiintyvyys kasvaa varhaisuusenglannin aikana viidestä lähes sataan! Meidän ei tarvitse nähdä ihan näin monta desimaalia, joten painellaan tätä Decrease Decimal -kuvaketta kunnes desimaaleja on enää yksi.

Havainnollistetaan kasvua tekemällä taulukon pohjalta viivadiagrammi. Valitaan taulukosta normalisoidut frekvenssit ja mennään Insert > Chart > Line. Huomaa, että nyt on hypätty Excelin työkalupalkissa Chart Design -välilehdelle. Muokataan otsikkoa ja annetaan kuvalle nimeksi "has". Lisätään X-akselille aikakaudet: painetaan Select Data > Horizontal axis labels, valitaan kausisolut ja painetaan OK. Lisätään akseleille myös otsikot: painetaan Add Chart Element > Axis Titles > Primary Horizontal ja kirjoitetaan "Aikakausi", sitten valitaan Primary Vertical ja kirjoitetaan "Normalisoitu frekvenssi / 100 000 sanaa". Voidaan myös muokata tuota Y-akselia niin, että se ei turhaan mene yli sadan: tuplaklikataan akselia ja kirjoitetaan Maximum-kohtaan 100. Kuvasta näkyy, että kasvu kiihtyy kahden jälkimmäisen kauden välillä.

Has-verbin normalisoitu frekvenssi kasvaa siis selvästi, mutta onko tämä *has*-muodon nousu tilastollisesti merkitsevä? Sitä pohditaan seuraavaksi.