VIDEO 3d: TILASTOLLINEN MERKITSEVYYS (TANJA SÄILY)
TRANSKRIPTIO

Tilastollinen merkitsevyys auttaa arvioimaan tuloksien luotettavuutta eli sitä, miten todennäköistä on, että havaittu muutos tai muu ilmiö on todellinen eikä pelkkää satunnaisvaihtelua. Kaius Sinnemäki kertoo siitä lisää omassa metodipankkiosiossaan. Merkitsevyyden mittaamiseen on monta eri tapaa. Monissa tutkimuksissa lasketaan merkitsevyyttä pelkkien normalisoitujen frekvenssien perusteella. Tässä esittelen sellaisen tavan, joka ottaa huomioon dispersion, eli sen, miten paljon sanaa käytetään eri teksteissä. Jos sana on jollakin aikakaudella yleisempi kuin toisella, mutta kaikki esiintymät ovat samasta tekstistä, tulos ei ole yhtä luotettava kuin silloin, jos esiintymät jakautuvat tasaisemmin tekstien välille.

Meillä on siis havaintojen perusteella syntynyt hypoteesi, että *has*-muotoa käytetään yhä enemmän varhaisuusenglannin aikana Helsinki Corpuksessa, niin että sen käyttö on erilaista viimeisellä kaudella kuin sitä edeltävällä kaudella. Sitten muotoillaan niin sanottu nollahypoteesi, johon ei oikeasti uskota ja joka halutaan kumota. Tässä tapauksessa nollahypoteesi voidaan muotoilla vaikka näin: se, esiintyykö *has*-muotoa tekstissä, ei riipu siitä, kummalla kaudella teksti on kirjoitettu. Tilastollinen merkitsevyys mittaa käytännössä sitä, onko meillä riittävästi evidenssiä nollahypoteesin kumoamiseen.

Tällaiseen hypoteesitestaukseen tarvitaan tieto siitä, montako tekstiä kullakin aikakaudella on ja monessako niistä esiintyy tai ei esiinny *has*-muotoa. Tekstien kokonaislukumäärät löytyvät metodipankista lataamastamme Excel-taulukosta, mutta monestako eri tekstistä löytyy *has*-muoto? Palataan has-tiedoston toiselle välilehdelle käsittelemään aineistoa.

Tässä meillä on kyllä lista kaikista teksteistä, joissa *has*-muotoa esiintyy, mutta tekstin nimi toistuu joka kerta kun siitä löytyy uusi esiintymä. Excelin avulla listan voi suodattaa niin, että kukin teksti esiintyy vain kerran per variantti ja aikakausi. Kopioidaan ensin näitten kolmen sarakkeen otsikot uuteen paikkaan. Sitten valitaan sarakkeet ja mennään Data > Advanced Filter. Valitaan Copy to another location ja Unique records only. Copy to -kohtaan valitaan uudet sarakkeet ja painetaan OK. Nyt voidaan analysoida näitä sarakkeita toisella Pivot-taulukolla. Valitaan sarakkeet ja mennään Data > Summarise with Pivot Table > OK. Raahataan riveille aikakausi ja sarakkeille variantti, ja lasketaan tekstien lukumäärää. Suodatetaan tyhjät pois ja siitäpä nähdään, monessako eri tekstissä *has*-muotoa esiintyy.

Kopioidaan tälle välilehdelle metodipankin Excelistä aikakausien tekstimäärät. Tämän jälkeen kopioidaan Pivot-taulukosta niitten tekstien määrä, joista löytyy *has*, taas niin että solut linkitetään

toisiinsa. Sitten halutaan laskea ne tekstit, joissa *has*-muotoa ei esiinny. Se onnistuu helpolla vähennyslaskulla: kaikki tekstit miinus *has*-tekstit. Aloitetaan taas kirjoittamalla soluun yhtäsuuruusmerkki. Sitten valitaan aikakauden tekstimäärä, lisätään viiva miinusmerkiksi, valitaan *has*-tekstien määrä ja painetaan enteriä. Täytetään tämä laskutoimitus muillekin riveille vetämällä solun oikeasta alakulmasta.

Nyt voidaan aloittaa hypoteesitestaus. Käytetään Fisherin tarkkaa testiä, johon on olemassa nettilaskuri GraphPad-sivustolla. Tässä vertaillaan aina kahta aikakautta keskenään. Vertaillaan ensin kahta ensimmäistä aikakautta. Nollahypoteesi, jota yritetään kumota, on tämä: *has*-muodon esiintyminen tekstissä ei riipu siitä, kummalla kaudella teksti on kirjoitettu. Ryhmä 1 on 1500–1570, ryhmä 2 on 1570–1640. Tulos 1 on *has*-tekstien lukumäärä ja tulos 2 on hasittömien tekstien lukumäärä. Ensimmäisellä kaudella tekstien määrät ovat 3 ja 24, toisella 6 ja 22. Painetaan "Calculate". Näitten kahden aikakauden välinen ero ei ole tilastollisesti merkitsevä. P-arvo on noin 0,47, mikä tarkoittaa sitä, että jos nollahypoteesi olisi totta, todennäköisyys saada vähintään näin iso ero aikakausien välille olisi 47 prosenttia, mikä on varsin suuri todennäköisyys. Niinpä nollahypoteesia ei tässä tapauksessa voida kumota. Tilastollisen merkitsevyyden rajana pidetään yleensä P-arvoa 0,05, mikä tarkoittaa vain viiden prosentin todennäköisyyttä saada tällainen ero aikakausien välille sattumalta.

Vertaillaan sitten kahta jälkimmäistä aikakautta. Nyt ryhmä 1 on 1570–1640 ja ryhmä 2 on 1640–1710. Tekstien lukumäärät ovat 6, 22, 19 ja 7. Lasketaan taas merkitsevyys. Laskurin mukaan P-arvo on 0,0003, mikä on erittäin merkitsevä. Se tarkoittaa sitä, että tämän testin mukaan on vain 0,03 prosentin todennäköisyys saada näin iso ero aikakausien välille pelkästään sattumalta. Näitten kausien välillä on siis tilastollisesti merkitsevä ero *has*-muodon käytössä.