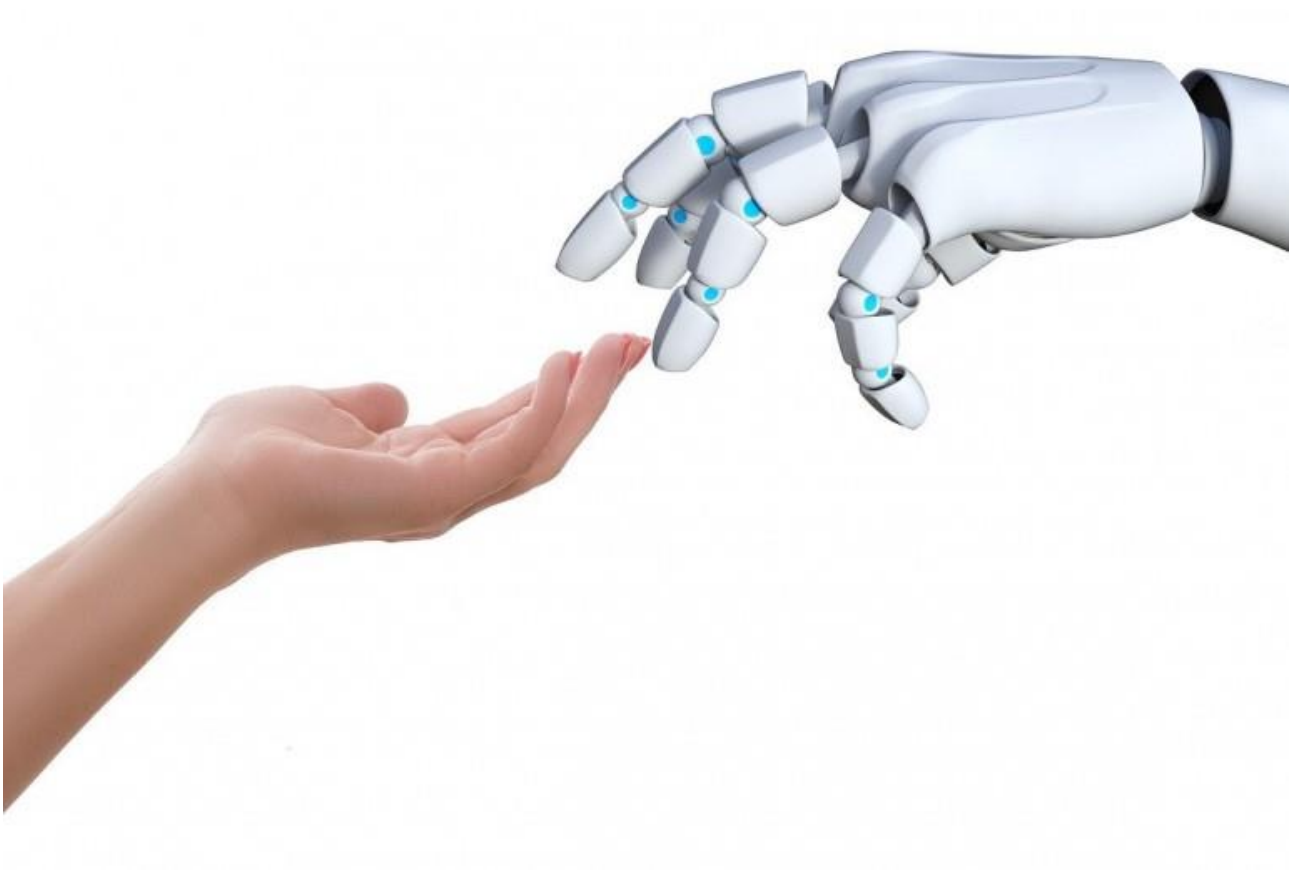

Kuva: Pete Linforth / Pixabay

19.11.2019, 15:55

TULE APUUN, ANNIF!

TIINA TOLONEN

Sisällönkuvailun on todettu olevan ihmiselle vaikeaa. Ihminen on subjektiivinen olento, joka näyttäytyy myös kuvailutyössä: kun kaksi eri ihmistä kuvailee saman dokumentin, vain noin 1/3 aiheista on samoja. Oman lisänsä tähän tuo käsitteiden valtava määrä sekä sanastojen muutokset. Joka vuosi kieleemme putkahtaa uusia termejä ja vanhoja määritellään uudelleen. Kuvailijan avuksi on Kansalliskirjastossa kehitetty tekstin perusteella sisällönkuvailua tuottava työkalu Annif, joka tulevaisuudessa on käytössä myös julkaisuarkisto Theseuksessa.

MÄÄRÄ VS. LAATU

Theseuksen asiasanalistauksessa on tällä hetkellä lähes 210 000 sanaa. Osa näistä on toki tuplia, sillä Theseus listaa saman termin kahteen kertaan, jos toinen on kirjoitettu isolla ja toinen pienellä alkukirjaimella. Oman lisänsä sanojen määrään tuovat kirjoitusvirheet. Jo pikaisella tutkimisella voi havaita, että monet asiasanalistan sanat esiintyvät siinä vain kertaalleen. Nämä ovat usein opiskelijan syöttövaiheessa kehittämiä avainsanoja, joita siis vanhalla syöttölomakkeella tallennettiin. Uudella syöttölomakkeella on ainoastaan asiasanakenttä. Ennen maaliskuuta 2019 tallennetuissa opinnäytetöissä on siis paikoitellen hyvin runsasta kuvailua, opiskelijan syöttämien avainsanojen lisäksi myös kirjastohenkilökunnan lisäämät asiasanat. Asiasanottamisesta on kuitenkin luovuttu kirjastoissa, ja kuvailu tapahtuu ainoastaan opiskelijan toimesta.

Esimerkki Theseuksen asiasanalistauksesta.

MIKÄ ANNIF? MIKSI ANNIF?

Annif on Finnaan perustuva sisällönkuvailupalvelu, jolle voi tarjota tekstejä tällä hetkellä suomeksi, ruotsiksi sekä englanniksi. Se perustuu kieliteknologiaan ja koneoppimiseen ja soveltuu parhaiten asiatekstille. Annif oppii tunnistamaan aiheita aineistosta. Kuten alussa viittasinkin, sisällönkuvailun on todettu olevan ihmiselle vaikeaa, vähintään haastavaa. Opiskelijalle voi jo pelkkä termi asiasana kuulostaa oudolta ja joskus tuntuu, että niiden valitseminen omalle opinnäytetyölleen on hankalampaa kuin itsensä työn kirjoittaminen. Se saattaa olla myös paljon aikaa vievä vaihe.

Annif-työkalun käytöstä asiasanottamisen helpottamiseksi on saatu kokemuksia Jyväskylän yliopistosta, jossa se on otettu käyttöön JYX-julkaisuarkistoon graduja tallennettaessa. Työkalua hyödynnettäessä gradun tallennus julkaisuarkistoon alkaakin ns. lopusta, sillä ensimmäisenä tallennetaan PDF-tiedosto, josta Annif tekee opiskelijalle asiasanaehdotuksia. Opiskelija valitsee niistä työhönsä sopivat, hänellä on mahdollisuus myös lisätä omia asiasanojaan vapaatekstikenttään. Informaatikko tarkistaa asiasanoituksen ennen gradun julkaisua.

Annifin käyttöä testattiin ensin Avoimen tieteen keskuksessa sisäisesti ja todettiin, ettei sen tuolloinen versio toiminut toivotulla tavalla. Kun uudempi versio sitten saatiin käyttöön voitiin todeta, että Annif on kehittynyt ja sen myötä noin 85 % opiskelijoista kelpuutti ainakin yhden Annifin tekemän asiasanaehdotuksen, kun se aiemmin oli 65–70 % luokkaa. Yli kolmasosa opiskelijoista valitsi Annifin tekemistä ehdotuksista yli puolet, osalle kelpasivat kaikki ehdotukset. Seuraavaksi Annifia ryhdytään kokeilemaan Avoimen tieteen keskuksessa uusien väitöskirjojen kanssa sekä myös museoaineistolla, jolle Annif saattaa antaa täysin uudenlaisen näkökulman.

ANNIFIN TULEVAISUUS JA THESEUS

Annifia siis kehitetään edelleen. Lähitulevaisuuden tavoitteina on parantaa kuvailun laatua testaamalla ja arvioimalla algoritmeja yhdessä CSC:n kanssa. Kehittämistä ovat vieneet eteenpäin myös Kirjastoverkkopäivillä vuosina 2017 ja 2019 järjestetyt työpajat, joista tämänsyksyinen **Aaveita koneessa** keräsikin mukavasti dataa esimerkkiasiakirjoille tehtyjen sisällönkuvailujen arvioinnin ja pisteyttämisen kautta. Työpaja antoi ajattelemisen aihetta myös minulle, joka olen sisällönkuvailutyötä tehnyt yli 20 vuoden ajan. Paikoitellen koneellisesti tai koneavusteisesti tehtyä kuvailua ei pystynyt lainkaan erottamaan ihmistyönä tehdystä, paikoitellen taas koneellisesti tehty kuvailu kompastui juuri siihen, että koneelta puuttui ihmisen kokemusperäinen tieto. Tai no, voihan Jeesusta kuvailla myös ainoana lapsena.

Lähitulevaisuuden tavoitteisiin kuuluu myös työkalun käyttöönotto Kansalliskirjaston järjestelmissä, joka tarkoittaa DSpace-pohjaisia julkaisuarkistoja sekä E-vapaakappaleiden vastaanottoa ja käsittelyä. Tätä kautta Annif on siis tulossa myös Theseukseen. Alustavan suunnitelman mukaan vuoden 2020 ensimmäisellä neljänneksellä toteutetaan suunnitteluvaihe ja toisella neljänneksellä suoritetaan pilotointi jonkin syöttövolyymiltaan pienen julkaisuarkiston kanssa. Tuotantovaiheeseen voitaisiin päästä pilotin jälkeen loppuvuodesta 2020 edellyttäen että

käytettävyys ja mahdollisesti esille tulevat ongelmakohdat on saatu ratkottua.

Kuten menneiltä vuosilta muistetaan, Theseus on julkaisuarkisto, jonka suuren koon ja syöttövolyymin vuoksi mahdolliset ongelmat yleensä moninkertaistuvat. Pienemmissä julkaisuarkistoissa monet kehitysvaiheet ovat menneet jouhevasti läpi, kun Theseuksen kanssa tehdään töitä hartiavoimin. JYXin hyvä esimerkki luo odotuksia myös Theseuksen suhteen, toivottavasti pääsemme ensi vuonna tätä kokeilemaan.

TIETOA KIRJOITTAJASTA:

Tiina Tolonen palvelupäällikkö, Oulun ammattikorkeakoulu