LOUHOKSELLA

Ihminen menee vaikka läpi harmaan kiven. Jos ei muuten, niin louhimalla. Lajiamme luonnehtii outo vimma murskata maaperää sekä rakennella louhoskivistä monenmoisia muureja vihollisten varalle ja temppeleitä itsensä jumalaksi korottaneille hallitsijoille.

Tunnelin louhiminen kallioon tietä varten, tienlouhinta, on jo arkipäivää. Vakaasti on kehitetty myös tekniikoita, jotka mahdollistavat uudenlaista harmaan kiven läpi menemistä – nimittäin tiedonlouhintaa.

Merriam-Webster-sanakirjan mukaan *data mining* -termiä käytettiin ensimmäisen kerran vuonna 1969 nykyisessä tietoteknisessä merkityksessä. Tieteen termipankki kiteyttää, että *tiedonlouhinta* viittaa joukkoon menetelmiä, joilla pyritään oleellisen informaation löytämiseen suurista tietomassoista.

Kielentutkijana olen jonkin verran perehtynyt laajojenkin ”tietomassojen” eli kieliaineistojen eli korpusten analysointiin. Jo nuorena opiskelijana 1980-luvun puolessavälissä pääsin kurkkaamaan Oulun yliopiston humanistisen tiedekunnan tietokonehuoneeseen. Iso masiina siellä ruksutti ja tulosti kilometrin mittaista lomaketta. Ihmeellistä! Ja Oulun korpus, mikä mahtiaineisto, yli 400 000 sanaa!

Kehitys on kehittynyt. Vuosituhannen vaihteessa analysoin kollegoideni kanssa yli 20 000 000 sanan korpusta. Nykyään internetistä on koottavissa miljardien sanojen aineistoja. Louhittavaa riittää!

Tunnustan, että olen pudonnut kehityksen kelkasta. Oikeastaan hyppäsin itse pois kyydistä siinä vaiheessa, kun kävi ilmeiseksi, että automaattiset analysaattorit eivät pystyneet tuomaan tyydyttäviä vastauksia kysymyksiini: Miten monitasoisia ja -ulotteisia merkityksiä synnytetään kokonaisissa teksteissä? Miten tekstilaji vaikuttaa kielellisten valintojen merkityksiin?

Suhtaudun yhä hieman varauksellisesti raportteihin, joiden sanotaan perustuvan tekstien koneelliseen analyysiin. Tällainen on esimerkiksi valtioneuvoston kanslian tilaama selvitys, joka liittyy koronarajoitusten purkamista koskevaan kansalaiskyselyyn. Kyselyn avovastausten analyysi tilattiin eräältä ”kulttuurin analytiikkayritykseltä”. Yritys tutki yli 2 000 vastausta, joissa on yhteensä yli 170 000 sanaa.

Louhituksi saatiin muun muassa sellainen tieto, että palautteiden sävy on ”useimmiten kriittinen ja negatiivinen” ja että ”perustelut ovat pääosin melko rakentavia”. Analyysi on siis sekä laadullista että määrällistä. Kysymyksiä herättää, kuinka syvälle koneellisessa analyysissa on päästy vastausten sävyihin ja vaikkapa mahdolliseen sarkasmiin tai ironiaan, vallankin kun yritys kertoo, että analyysissa on suodatettu pois ”täytesanat”. Kielentutkijana olen vankasti sillä kannalla, että teksteissä ei ole täytesanoja – jokaisella valinnalla on tehtävänsä merkitysten tuottamisessa.

Suhtaudun kuitenkin erittäin myönteisesti isojen aineistojen koneelliseen analyysiin ja valtavankin datan moninaiseen tiedonlouhintaan. Olen varma, että ihminen pystyy opettamaan koneelle kaiken tarvittavan, myös täytesanojen merkityksellisyyden.

Ja ne tekstejäkin louhivat koneet, tervetuloa Suomeen! Lehdissä kerrotaan milloin minkäkin yrityksen datakeskuksen rakentamisesta monttuun tai luolaan, joka on louhittu suomalaiseen peruskallioon.

VESA HEIKKINEN
Kirjoittaja on suomen kielen dosentti ja tietokirjailija.
Twitter: @tosentti