Risto Turunen

Humanisti, kurkista makroskooppiin

Omalla kohdallani digitaalisen humanismin aurinko nousi Helsinki Digital Humanities Hackathonissa toukokuussa 2015. Hackathonin perusidea on yksinkertainen: humanistit ja tietojenkäsittelytieteilijät hikoilevat viikon yhteisen ongelman äärellä. Humanistien vastuulla on merkityksellisten tutkimuskysymysten keksiminen, kun taas ohjelmoijat rakentavat algoritmeja, joiden avulla kysymyksiin saadaan vastauksia. Ensimmäisessä hackathonissa ujutin historian väitöskirjani tutkimuskysymyksen oman ryhmäni mietittäväksi: mikä erottaa 1900-luvun alun sosialismin poliittisen kielen aikakauden muista kilpailevista käsitejärjestelmistä? Kansalliskirjasto antoi käyttöömme todellisen aarteen: kaikki suomalaiset sanomalehdet koneluettavassa muodossa vuosilta 1771–1910, miljardien sanojen ”kulttuurisen arkin”.

Mutta kuinka operationalisoida sosialismin kieli? Päädyimme lopulta melko primitiiviseen ratkaisuun eli laskimme, mitkä sanat toistuvat sosialistilehdissä useammin kuin koko lehdistössä. Parin päivän puurtamisen jälkeen koodarimme ilmoitti ensimmäisistä tuloksista. Hieman kohmeloisilla silmillä – sosiaaliset iltamenot kuuluvat tapahtumaan – tutkin läppärin näytöllä vilistäviä sanoja ja numeroita. Valistunut arvaaja olisi voinut ennustaa suuren osan sosialisteille tärkeistä sanoista (”kapitalismi”, ”porwaristo”, ”sorto”), mutta tuloksissa oli myös virkistäviä yllätyksiä (”piru, ”ryssä”, ”naiset”). Tietokone kykeni näyttämään sanomalehdistä jotakin, mitä ihmissilmän on lähes mahdotonta huomata.[1]

Suomalaisista historioitsijoista Tampereen yliopiston Viljo Rasila hyödynsi tietokonetta jo 1960-luvulla tutkiessaan sisällissodan syntyä monimuuttujamenetelmien avulla[2]. Rasila suositteli tietokoneen käyttöä silloin, ”kun on kysymys niin laajasta tavalla tai toisella mitattavasta aineistosta, että sen täydellinen hallitseminen on ihmisen omalle muistikyvylle mahdotonta, taikka jos vertailut ovat niin monimutkaisia, että ihminen ei kykene niitä suorittamaan”[3]. Rasilan neuvossa on edelleen järkeä: laskennallisten menetelmien mahdollinen hyöty riippuu tutkimusasetelmasta. Jos kartoittaa 1500-luvulla eläneen myllärin maailmankuvaa fragmentaaristen lähteiden avulla, tietokoneen kyvyistä ei liene suurta apua[4]. Esimerkiksi digitaalisten lehtien sisältöä kvantifioimalla tutkimani ilmiö koostuu puolestaan sadoistatuhansista sosialisteista ja heidän miljoonista sanoistaan, joten sitä

”’Terveeseen järkeen’ ei kannata luottaa liikaa menneisyyden tulkinnassa.”

on vaikea ottaa haltuun pelkästään ihmisen omilla kognitiivisilla kyvyillä.

Tietokone laskee tarkemmin kuin paraskaan matemaatikko. Se kykenee muokkaamaan tuhattakin tutkimusavustajaa nopeammin kaikki suomenkieliset sanomalehdet taulukoksi, joka näyttää, mitkä sanat toistuvat erityisen usein sosialistilehdissä. Toisin sanoen tietokone on tarkka, nopea ja väsymätön jalostaessaan raakadataa alkeelliseksi informaatioksi. Mitä enemmän tutkijalla on käytössään laadukasta informaatiota, sitä parempia tulkintoja – ja samalla humanistista tietoa – hän pystyy tuottamaan. (Toisaalta: jos alkuperäinen data on huonolaatuista, virheet kertaantuvat myöhemmin!)

”Terveeseen järkeen” ei kannata luottaa liikaa menneisyyden tulkinnassa. Psykologi Daniel Kahnemanin tunnetuksi tekemä WYSIATI – *What You See Is All There Is* – vaanii myös historiantutkijoita[5]. Kukapa meistä ei olisi joskus tehnyt yli-itsevarmoja johtopäätöksiä vähäisen ja huonolaatuisen informaation pohjalta. Historiantutkimuksessa huono laatu voi tarkoittaa esimerkiksi informaation sirpaleisuutta. Laajojen yhtenäisten aineistojen laskennallinen analyysi on hyvää lääkettä intuitiivisen ajattelun systemaattisia virheitä vastaan. Tämä ei tarkoita perinteisistä humanistisista menetelmistä luopumista: kaikki etälukemiseen perustuvat löydöt pitää aina varmistaa, kontekstoida ja tulkita lähilukemalla.

Toki digitaaliseen humanismiin ja digitaaliseen historiaan sen alalajina liittyy monia ratkaisemattomia ongelmia. Koneellisen tekstintunnistuksen (*optical character recognition*) kehnous vaivaa niin suomalaista sanomalehtikorpusta kuin monia muitakin mielenkiintoisia historiallisia digiaineistoja. Esimerkkinä mainittakoon, että suomenkielisissä sanomalehdissä esiintyy vuosina 1820–1910 yli 30 000 ”pillua”, joista tuskin yksikään on aito[6]. Lähilukemisen perusteella tietokone on tulkinnut sanan ”pikku” toistuvasti ”pilluksi”. Vaatii taitoa ja kärsivällisyyttä arvioida, mitkä löydöistä ovat oikeita signaaleja, kun korpuksessa on paljon ylimääräistä kohinaa – tätä voisi verrata rikkinäisen radion kuunteluun.

Toisena ongelmana on esitetty, että digitaalisuuden paine saattaa ohjata tutkijoita niiden harvojen aineistojen ääreen, jotka toistaiseksi on digitoitu. Tällöin fyysiset arkistokäynnit vähenevät, tutkimusaiheet vinoutuvat ja humanistisen tutkimuksen monimuotoisuus vaarantuu. Ongelmaa ei kuitenkaan ratkaista vanhojen kunnon arkistopäivien nostalgisella muistelulla ja digitaalisuuden paheksumisella, vaan historiantutkijoiden tulisi vaikuttaa aktiivisesti siihen, millaisia aineistoja tutkijoilla on tulevaisuudessa käytössään.[7]

Suomalaisen historiantutkimuksen kontekstissa digitaalisuuden ylivaltaa suurempi ongelma saattaa tällä hetkellä olla se, että tutkijat eivät käytä digitaalisia aineistoja

riittävästi – ja silloin kun käyttävät, aineistoja katsotaan yksiulotteisesti vanhojen silmälasien läpi. Digitaalisia lehtiä on kyllä hyödynnetty onnistuneissa tutkimuksissa mutta lähes poikkeuksetta samalla tavalla: tutkija etsii omaan aiheeseensa liittyvät tekstit sopivilla hakusanoilla Kansalliskirjaston Digi-palvelussa. Sanahakujen tekeminen koko lehdistöön on pienimuotoinen tutkimuksellinen läpimurto, sillä sen avulla löydetään aikaisempaa suurempi osa tutkimuskysymyksille relevanteista primäärilähteistä. Digitaalisuus mahdollistaa kuitenkin myös subjektiivisia sanahakuja syvällisempiä lähestymistapoja. Yksi lupaava esimerkki tästä on Suomen Akatemian rahoittama COMHIS-projekti, jossa tutkitaan muun muassa kulttuurin viraalisuutta kvantifioimalla tekstien uudelleenjulkaisua sanoma- ja aikakauslehdistössä[8]. Tällaisissa tutkimuksissa korostuu inhimillisten ennakkooletusten sijaan tietokoneen toimijuus: kone ei etsi *tiettyjä* lehdissä kiertäviä tekstejä vaan *kaikki* uudelleenjulkaistut tekstit. Näin tutkija voi parhaassa tapauksessa tavoittaa ilmiöitä, joiden olemassaoloa hän ei kyennyt ennalta edes kuvittelemaan.

Digitaalisia menetelmiä kekseliäästi soveltavat projektit saattavat onnistuessaan ratkaista kolmannen ongelman eli digitaaliseen historiantutkimukseen innostavien klassikkoteosten puutteen. Jos digitaalinen historia haluaa vakiinnuttaa asemansa varteenotettavana suuntauksena, se tarvitsee tuoreita tuloksia. Digihumanisti Matthew Jockersin kokemusten mukaan tuloksilla, jotka varmentavat jo olemassa olevia hypoteeseja, ei ihmistieteilijöiden keskuudessa jostakin kumman syystä nähdä suurta arvoa toisin kuin luonnontieteissä, joissa lisätodisteet esimerkiksi evoluutioteorian puolesta otetaan tyytyväisenä vastaan[9]. Kun julkaisukynnyksen ylittäviä tuloksia lopulta saadaan aikaan, ne pitäisi jaksaa asettaa vuoropuheluun aiemman tutkimuksen kanssa sekä esittää vetävästi ja selkeästi. DH-artikkeleista tutut värikkäät, esteettiset, joskus jopa hypnoottiset kuviot saattavat hämmentää lukijaa, vaikka visualisaation tarkoituksena pitäisi kai olla asioiden yksinkertaistaminen.

Olen käynyt hackathonissa kolmena keväänä peräkkäin maksamassa älyllistä velkaani digitaalisten humanistien yhteisölle. Ennen ensimmäistä reissuani olin lähes varma, että sosialismin erottaa kilpailijoistaan ennen kaikkea hullu usko tulevaisuuteen. Gradua tehdessäni olin lähilukenut yhtä Tampereen tehdassosialistien käsinkirjoitettua lehteä, jossa tulevaisuudesta puhuttiin runsaasti. Alustavat tulokset eivät kuitenkaan tukeneet ennakkokäsitystäni, sillä niiden mukaan sosialistilehdissä korostuvat nimenomaan nykyisyyteen viittaavat käsitteet eli ”nyky*”-alkuinen sanasto. Tällaisia määrällisiä eroja usein toistuvissa sanoissa on vaikea havaita perinteisellä lähilukemisella. Mutta mitä nämä erot aikasanoissa oikein merkitsevät? Ehkä tulevaisuusvisiot olivat tärkeitä myös 1900-luvun alun uskonnollisille ja porvarillisnationalistisille lehdille, vaikka niiden kuvittelema tulevaisuus erosikin laadullisesti punaisesta huomisesta. Ehkä sosialistien nykyisyyden korostaminen johtui siitä, että he näkivät ympärillään niin paljon epäkohtia. Aion ratkaista sosialistisen aikakäsityksen arvoituksen viimeistään väitöskirjani päätösluvussa.

Merkitysten luominen vie aikaa ja ”historia on hidas tiede”, kuten historian väitöskirjatutkijoilla on tapana vastata uteluihin tutkinnon valmistumisesta. Lisätään tähän vielä yksi adjektiivi: historia on myös empiirinen tiede. Menneisyydestä jää jälkiä, joita on mahdollista tutkia vaihtoehtoisin metodein. Viljo Rasila käynnisti aikoinaan jonkinlaisen kvantitatiivisen käänteen suomalaisessa historiantutkimuksessa, mutta etenkin 1990-luvulla ajan henki voimisti akateemisia virtauksia, joissa painotettiin historiallisten toimijoiden yksilöllisyyttä ja ainutlaatuisuutta (”uusi sosiaalihistoria”, ”mikrohistoria”, ”psykohistoria”). Kenties 2010-luvun digitaalinen humanismi hackathoneineen tarkoittaa taas liikettä yksilöistä rakenteisiin sekä lyhyestä pitkään aikaväliin.

Metaforisesti: mikroskoopin ohella historiantutkijat voisivat vaihtelun vuoksi tarttua myös ”makroskooppiin”. Mikroskoopilla havaitaan paremmin erittäin pieniä ilmiöitä, kun taas makroskooppi auttaa ymmärtämään erittäin suuria ja kompleksisia ilmiöitä. Makroskooppi vähentää tietoisesti ilmiön monimutkaisuutta, jotta tavoitettaisiin sen aiemmin pimentoon jääneitä ominaisuuksia.[10] Tammerkoskessa on virrannut paljon vettä sitten Rasilan päivien: jokaisen nykyopiskelijan läppärissä on valtavasti enemmän laskentatehoa kuin 1960-luvun supertietokoneessa, ja uusia houkuttelevia digiaineistoja syntyy kaiken aikaa. Nyt on hyvä hetki kurkistaa makroskooppiin ja tehdä digihistoriaa.

Viitteet & Kirjallisuus

1 Sosialismin avainsanoja voi tutkia osoitteessa: github.com/dhh15/fnewspapers/tree/master/results

2 Viljo Rasila, *Kansalaissodan sosiaalinen tausta*. Tammi, Helsinki 1968.

3 Viljo Rasila, Ajankohtaisia tutkimuskysymyksiä – Tietokone historiantutkimuksessa. *HAik* 2/1967, 146.

4 Viittaan mikrohistorian klassikkoon eli Carlo Ginzburgin teokseen *Juusto ja madot – 1500-luvun myllärin maailmankuva* (Il formaggio e i vermi, 1976). Suom. Aulikki Vuola. Gaudeamus, Helsinki 2007.

5 Daniel Kahneman, *Thinking, Fast and Slow*. Farrar, Straus & Giroux, New York 2011.

6 Perustuu hakuun ”pillu” suomenkielisissä sanomalehdissä vuosina 1820–1910 osoitteessa digi.kansalliskirjasto.fi/sanomalehti/search

7 Jo Guldi & David Armitage, *The History Manifesto*. Cambridge University Press, Cambridge 2014, 113.

8 Ks. esim. Aleksi Vesanto, Asko Nivala, Heli Rantala, Tapio Salakoski, Hannu Salmi & Filip Ginter, Applying BLAST to Text Reuse Detection in Finnish Newspapers and Journals, 1771–1910. *Proceedings of the 21st Nordic Conference of Computational Linguistics.* Göteborg 23.–24.5.2017, 54–58. Verkossa: www.ep.liu.se/ecp/133/010/ecp17133010.pdf

9 Ks. Matthew L. Jockers, *Text Analysis with R for Students of Literature*. Springer, New York 2014, vii–viii.

10 Shawn Graham, Ian Milligan & Scott Weingart, The Joys of Big Data for Historians. Teoksessa *The Historian's Macroscope – Big Digital History*. Imperial College Press. Open Draft Version, 2013. Verkossa: themacroscope.org