”Big datan” haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät

Esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935–2015

Semi Purhonen & Arho Toikka

VERTAISARVIOITU
KOLLEGIALT GRANSKAD
PEER-REVIEWED
www.tsv.fi/tunnus

Abstrakti

Artikkeli on keskusteluavaus, jossa pohditaan digitalisoitumisen, ”big datan” ja uusien laskennallisten analyysimenetelmien merkitystä ja antia sosiologialle. Lähtökohtana on, että vaikka usein esitetty väite big datasta rinnakkaisilmiöineen sosiaalitutkimuksen kannalta jopa vallankumouksellisena voi olla liioiteltu, big datan haaste ja mahdollisuudet tulee ottaa vakavasti. Toisaalta, vaikka yleinen keskustelu big datan kokonaismerkityksestä sosiologialle on sinänsä tärkeää, keskustelu merkityksestä on hedelmällisempää, kun ilmiökenttää arvioidaan rajatummin. Lähempään tarkasteluun on artikkelissa valittu uudet laskennalliset tekstiaineistojen analyysimenetelmät ja esimerkkinä niistä aihemallit (*topic models*), jotka nähdään sosiologisen tutkimuksen kannalta lupaavina uusina välineinä monenlaisten ja -kokoisten tekstiaineistojen analysoimiseksi. Artikkeli esittelee aihemallit menetelmänä ja osoittaa Suomen presidenttien uudenvuodenpuheaineistoa koskevalla aihemallisovelluksella, että menetelmää voidaan hyödyntää melko rajattujenkin tekstiaineistojen analysoimisessa. Uudet laskennalliset tekstidatan analyysimenetelmät eivät lähtökohtaisesti korvaa vaan täydentävät perinteistä ”kvalitatiivista” luku- ja tulkintatapaa. Big data ja uudet laskennalliset menetelmät ovat silti enenevässä määrin relevantteja myös aivan ”tavallisen” kokoisia ja kvalitatiivisiksi perinteisesti miellettyjä aineistoja analysoivien sosiologien kannalta.

ASIASANAT: Aihemallit, big data, digitaaliset ihmistieteet, laskennallinen sosiologia, Suomi, tasavallan presidentin uudenvuodenpuheet, tekstiaineistot

Johdanto

Viime vuosina lukemattomat eri alojen kommentaattorit, analyytikot ja tutkijat ovat esittäneet, että digitalisoituminen, erilaisten aineistojen – ”datan”, kuten puheenparsi suomeksikin kuuluu – lisääntyminen ja tuon datan analyysivälineiden kehitys merkitsevät tärkeää, jopa vallankumouksellista yhteiskunnallista murrosta; vain akatee-

misuuden aste ja sävy, jolla murrosta kuvaillaan, vaihtelee populaarista ja avoimen entusiastisesta (Mayer-Schönberger & Cukier 2013; Pentland 2014) analyyttisempaan ja skeptisempään (boyd & Crawford 2012; Kitchin 2014).

Murrospuheessa on kaksi puolta. Ensinnäkin nähdään, että murros koskee yhteiskunnallista todellisuutta, siis sosiologian tutkimuskohdetta. Melko laajan konsensuksen mukaan internetin, digitaalisen teknologian ja etenkin sosiaalisen median nousun tuottama datan määrä johtaa ”datavallankumoukseen”, jossa yhteiskunta muuttuu olennaisesti datan tullessa uudella tavalla liiketoiminnan, tutkimuksen, politiikan ja hallinnan sekä tavallisten ihmisten elämän keskiöön (Bail 2014; boyd & Crawford 2012; Dutton & Graham 2014; Kitchin 2014). Toiseksi kysymys on siitä, missä määrin tämä yhteiskunnallisen todellisuuden muutos merkitsee murrosta myös sosiologisen tiedonmuodostuksen ja tutkimuskäytäntöjen kannalta (Burrows & Savage 2014; Ruppert, Law & Savage 2013; Savage & Burrows 2007; 2009; Schroeder 2014; Tinati ym. 2014).

Tässä artikkelissa, joka on keskustelunavaus ja kutsu pohtimaan digitalisoitumisen, big datan ja uusien laskennallisten analyysimenetelmien merkitystä, tarkastellaan kysymystä ensi sijassa jälkimmäisestä, sosiologian tieteenalan, näkökulmasta. Artikkelin ensimmäisenä lähtökohtana on, että vaikka usein esitetty ajatus big datasta sosiaalitutkimuksen kannalta jopa vallankumouksellisena voi olla liioiteltu (big data -puheeseen kuuluvan teknologisen hybriksen voi nähdä jopa elimellisenä osana ilmiötä; ks. boyd & Crawford 2012), tulee sosiologian ottaa big datan haaste ja mahdollisuudet vakavasti. Jos big dataa ei ymmärretä kirjaimellisesti vain kokoon viittaavina ”isoina aineistoina”, vaan paremminkin erityisenä ja uusia menetelmiä kehittävänä tapana tarkastella ja hyödyntää moninaisia aineistoja, on nähdäksemme selvää, että big data merkitsee varteenotettavaa muutosta sosiologian kannalta. Toinen lähtökohta on, että vaikka yleisen tason keskustelu big datan kokonaismerkityksestä sosiologialle on sinänsä tärkeää ja tervetullutta, merkitystä koskeva keskustelu on kuitenkin hedelmällisempää, kun tarkastellaan ja arvioidaan ilmiökenttää konkreettisemmin ja yksityiskohtaisemmin, rajautuen vain johonkin osaan sitä.

Big dataan, digitalisoitumiseen ja uusiin laskennallisiin analyysimenetelmiin liittyvään kansainväliseen keskusteluun ovat osallistuneet viime vuosina tietojenkäsittelytieteen ja sen rinnakkaisalojen kiistämättömästä hallinnasta huolimatta enenevissä määrin myös historioitsijat ja humanistit (vrt. puhe ”digitaalisesta ihmistieteistä”, *digital humanities*; Gold 2012) sekä yhteiskuntatieteilijät (vrt. puhe ”laskennallisesta yhteiskuntatieteestä”, *computational social science*; Lazer ym. 2009). Kiinnostuksesta jälkimmäisten keskuudessa kertovat monien sosiologiankin kannalta keskeisten julkaisujen aihepiiriä käsitelleet erikoisnumerot (esim. *The Annals of the American Academy of Political and Social Science* 1/2015; *Poetics* 6/2013; *Sociology* 6/2012; *Theory and Society* 3–4/2014). Tietynlaisena institutionalisoitumisen merkkinä alalle on perustettu myös omia julkaisusarjoja (esim. *Big Data & Society* 2014 alkaen). Alalle luonteenomaisesti koko akateeminen keskustelu laahaa tosin kaupallisten tahojen perässä, sillä etenkin Googlen ja Facebookin kaltaisten jättiyritysten tutkimusosastot ovat toimineet kehityksen suunnannäyttäjinä, joskin niiden ja yhdysvaltalaisten huippuyliopistojen tutkimuksen välillä on ollut merkittävää yhteistyötä.

Suomalainen keskustelu aihepiiristä on ollut toistaiseksi vähäistä. Institutionaalisia heräämisen merkkejä ovat esimerkiksi Suomen Akatemian huhtikuussa 2015 julkistama digitaalisten ihmistieteiden akatemiaohjelma (2016–19), Helsingin yliopiston tutkijakollegiumin vuoden

2014 lopulla käynnistämät seminaarit sekä Aaltoyliopiston toimesta Helsingissä kesäkuussa 2015 järjestetty suuri kansainvälinen laskennallisen yhteiskuntatieteen konferenssi. *Sosiologiassa* tai muissa suomalaisen sosiologian keskeisissä julkaisuissa ei ole toistaiseksi käyty lainkaan keskustelua big datasta tai sen rinnakkaisilmiöistä. Tuore sosiologian johdantoteos, joka kertoo tarkoituksekseen ”johdattaa lukija 2000-luvun sosiologian olennaisten peruskysymysten äärelle” (Erola & Räsänen 2014, 241), mainitsee termin ”big data” ohimennen yhdellä sivulla käsittelemättä aihetta tarkemmin.

Tämän artikkelin tarkoituksena on herättää keskustelua big datasta myös suomalaisessa sosiologiassa. Koska katsomme, että keskustelua big datan ja sen mukanaan tuomien uusien menetelmien merkityksestä kannattaa käydä konkreettisesti, on artikkelissa valittu lähempään tarkasteluun uudet laskennalliset tekstiaineistojen analyysimenetelmät ja esimerkkinä niistä aihemallit. Internetin ja sosiaalisen median kehittyminen samoin kuin erilaisten digitalisoitujen arkistojen ja tietopankkien karttuminen on johtanut tekstiaineistojen määrän hurjaan kasvuun, mikä on tuonut uudet tekstidatan analyysimenetelmät big datan keskiöön (Bail 2014). Näemme nämä menetelmät sosiologisen tutkimuksen kannalta erityisen lupaavina uusina välineinä monenlaisten, yleensä ”kvalitatiivisiksi” miellettyjen, tekstiaineistojen analysoimiseksi. Pyrimme osoittamaan Suomen presidenttien uudenvuodenpuheaineistoa hyödyntävällä aihemallisovelluksella, että aihemalleja voidaan käyttää hyödyllisesti melko rajatunkin tekstiaineiston analysoimisen apuna. Ylipäätään pyrimme arvioimaan big datan ja etenkin tekstidatan laskennallisten menetelmien etuja ja rajoituksia sosiologian kannalta sekä argumentoimaan sellaisen maltillisen kannan puolesta, että vaikka big data tai sen mukanaan tuomat uudet menetelmät eivät mullistaisikaan sosiaalitieteitä, on kysymyksessä tärkeä ja varteenotettava kehityssuunta, josta sosiologia ei voi jäädä jälkijunaan.

Big data, uudet laskennalliset analyysimenetelmät ja sosiologia

Nykyistä datan määrää tai ”datatulvaa” – maailman dataistumista, datafikaatiota (Mayer-Schönberger & Cukier 2013, 73–97), jossa yhä useammalta elämänalueelta kerääntyvä tieto paisuu – on kuvattu seuraavasti: vuoden 2002 aikana kertyi enemmän dataa kuin koko aiemman ihmiskunnan historian aikana, ja vuoteen 2011 mennessä saman verran dataa kuin vuotta 2002 ennen kasautui jo joka toinen päivä (Bail 2014, 465). IBM raportoi vuonna 2012, että yli 90 % kaikesta maailman datasta sillä hetkellä oli luotu viimeisen kahden vuoden aikana (Kitchin 2014, 69). Dataa syntyy nykyisin radikaalisti enemmän kuin ennen ja kasvuvauhti vain jatkuu. Datan koko, tai se, että aineistoa on paljon, ei kuitenkaan tavoita big data -ilmiöstä kuin yhden puolen. Big datalla on monia ominaispiirteitä, jotka tekevät sen erityiseksi perinteisiin aineistoihin verrattuna (emt., 79). Big datan voikin ajatella olevan sateenvarjokäsite, joka kattaa niin aineistojen syntyyn ja tuottamiseen, säilyttämiseen ja jakamiseen kuin analyysiin ja jatkokäyttöön liittyvät ulottuvuudet. Tällöin big data -ilmiön piiriin voidaan laskea myös sen rinnalla kehittyneet uudet aineistojen analyysimenetelmät.

Termin big data käyttö yleistyi 2000-luvun ensimmäisen vuosikymmenen loppupuolella, ennen muuta amerikkalaisen *Wired*-lehden kaltaisten tietotekniikan uutuuksia ja kulttuurista merkitystä käsittelevien kaupallisten medioiden piirissä, lyödäkseen läpi laajemmassa mittakaavassa ja moninaisemmilla areenoilla 2010-luvun ensimmäisinä vuosina (Burrows & Savage 2014; Mayer-Schönberger & Cukier 2013, 1–18). Big datalle ei ole

olemassa yhtä yleisesti hyväksyttyä määritelmää (ks. esim. Dutcher 2014, joka on koonnut sisällöltään hyvinkin vaihtelevan 40 määritelmän listan).

Yleisin big datan määritelmä viittaa v-kirjaimilla alkaviin sanoihin, yleensä kolmeen (esim. Kitchin 2014). Alkuperäisen englanninkielisen version lisäksi tämä toimii myös suomeksi. Ensinnäkin big data on kooltaan *valtavaa* (*volume*), toiseksi nopeaa *vauhdiltaan* (*velocity*) sekä kolmanneksi muodoltaan *vaihtelevaa* (*variety*). Määrään liittyvän ilmeisen kriteerin lisäksi tarkoitetaan siis sitä, että data muuttuu nopeasti tai sitä syntyy (usein reaaliaikaisesti) vauhdilla lisää ja että data on moninaista, sisältäen sekä enemmän että vähemmän jäsentynyttä tai strukturoitua aineistoa. Strukturoitu data tarkoittaa binaarista tai kvantitatiivista dataa, joka kattaa suuren osan datatulvasta. Kuitenkin myös erilaisten tekstiaineistojen – kuten Twitter-viestien tai digitalisoitujen tekstiarkistojen (esim. *Google Books* tai sanoma- ja aikakauslehtien arkistot kuten *ProQuest*) – kasautuminen ja kerääntyminen on ennennäkemätöntä. Nämä valtavat tekstimassat ovat erityisen houkuttelevia kulttuurisosiologian kannalta, koska ne ovat ”luonnollisia” aineistoja (niiden olemassaolo ei ole ollut riippuvaista tutkijoiden interventiosta, toisin kuin esimerkiksi haastatteluissa) ja usein luonteeltaan ajallisen muutoksen analyysiin sopivia pitkittäisaineistoja. Siten ne antavat mahdollisuuden tutkia merkitysrakenteiden muutoksia *in situ* (emt., 467; myös Mohr & Bogdanov 2013). Tietojenkäsittelytieteessä on kehitetty uusia välineitä tällaisen ei-strukturoidun tekstidatan analyysia varten, mikä vaatii korkeaa laskentatehoa ja aiempaa kehittyneempiä algoritmeja (Bail 2014; Kitchin 2014; Mohr & Bogdanov 2013).

Mainitun kolmen ”v:n” lisäksi kirjallisuudessa on annettu big datalle lukuisia lisämääreitä. Näitä ovat esimerkiksi tyhjentävyys tai kaikenkattavuus (vrt. iskulauseet ”n = kaikki” ja ”*more is better*”; Mayer-Schönberger & Cukier 2013, 19–31), hienojakoinen resoluutio eli pyrkimys mahdollisimman detaljoituun informaatioon, indeksimäinen tunnistuskyky sekä datan relationaalisuus ja joustavuus (Kitchin 2014, 68). Tiettyä sosiologista houkuttelevuutta on boydin ja Crawfordin (2012, 663) määritelmässä, jonka mukaan big data on ”kulttuurinen, teknologinen ja tutkimuksellinen ilmiö”, joka on tulosta teknologian, analyysin ja mytologian välisestä vuorovaikutuksesta. Määritelmän teknologinen komponentti tarkoittaa ”laskentatehon ja algoritmitarkkuuden maksimointia pyrkimyksenä kerätä, analysoida, yhdistellä ja vertailla isoja aineistoja”; analyysi viittaa ”isoihin aineistoihin perustuvaan säännönmukaisuuksien tunnistamiseen taloudellisten, sosiaalisten, teknisten ja lainopillisten väitteiden esittämiseksi”; kun taas mytologia tarkoittaa määritelmässä ”laajalti levinnyttä uskomusta, että isot aineistot tarjoavat korkeamman tason älyä ja tietoa, joka voi luoda aiemmin mahdottomana pidetyn kaltaista tietämystä, jolla on totuuden, objektiivisuuden ja tarkkuuden aura” (emt., 663). boydin ja Crawfordin – molemmat akateemisia huippututkijoita, jotka artikkelin julkaisemisen aikaan työskentelivät Microsoftin tutkimusosastolla – määritelmän henkeä seuraten tekeekin mieli nostaa esiin vielä neljäs ”v”, joka määrittää vahvasti big data -ilmiötä ja keskustelua: arvo, hyöty ja viime kädessä *voitto* (*value*), joka usein motivoi big datan keruuta ja käyttöä.

Yhteiskuntatieteilijöiden piirissä reaktiot big dataan ovat olleet vaihtelevia. Big data on nähty yhteiskuntatieteissä yhtäältä houkuttelevana ja sisältävän huomattavaa potentiaalia, toisaalta uhkana ja negatiivisena kehityksenä. Uusien aineistomuotojen ja menetelmien houkuttelevuus on ymmärrettävää tilanteessa, jossa yhteiskuntatutkimusta luonnehtivat alati heikkenevät kyselyiden vastausprosentit ja huolet kvalitatiivisten tutkimusten edustavuudesta (Bail 2014, 466). Uhkana on nähty big datan kytkeytyminen mahdollisuuteen tarkkailla, valvoa ja hallita kuluttaja-kan-

salaisia entistä totaalisemmin ja uusilla tavoilla (Lyon 2014). Yritysten tavat käyttää laajoja asiakastietokantoja tutkimus- ja markkinointitarkoituksiin synnyttävätkin uusia eettisiä kysymyksiä, joita ei ole vielä ratkaistu (boyd & Crawford 2012). Joka tapauksessa lienee selvää, että big data merkitsee tärkeää trendiä ja muutosta myös yhteiskuntatieteille – ja että tuo muutos on muutos kohti laskennallisuutta ja siten hyvin spesifiä käsitystä tieteellisyydestä (Schroeder 2014).

Aiemmin data on ollut niukka hyödyke; sen kerääminen on ollut hankalaa, kallista ja vain tietyn spesialistiryhmän vaivalloisesti hallittavissa. Big data tuo tähän muutoksen, minkä voi nähdä syynä sen potentiaaliseen kriisivaikutukseen sosiologian kannalta (Burrows & Savage 2014; Savage & Burrows 2007). Teknologinen muutos – ennen muuta henkilökohtaisten tietokoneiden yleistyminen ja internetin nousu – ovat aiheuttaneet sen, että entistä useammat ihmiset entistä moninaisemmista taustoista voivat käyttää, tuottaa, jakaa ja organisoida dataa (boyd & Crawford 2012, 664). Savagen ja Burrowsin (2007) mukaan tämä merkitsee empiirisen sosiologian kriisiä, koska yhteiskuntatieteilijät ovat menettäneet etuoikeutensa dataan. Nykyisin se on lähes kaikkien riittävästi motivoituneiden kansalaisten saatavilla. Kun etuoikeus dataan häviää, häviää myös etuoikeus määritellä yhteiskunnallisesti merkityksellisen tiedon luonne (Burrows & Savage 2014, 5).

boyd ja Crawford (2012) vertaavat big datan aikaansaamaa muutosta tutkimukselle siihen, miten fordismi muutti tehdastuotannon perusteita 1900-luvun alkupuolella: kyse ei ollut vain liukuhihnoista ja tuotannon uudelleenjärjestelystä, vaan tehtailta käynnistyneet muutokset koskivat lopulta sitä, mikä oli ihmisten suhde työhönsä ja miten he ymmärsivät koko yhteiskunnan. Vastaavalla tavalla voidaan nähdä, että big data ei viittaa vain isoihin aineistoihin ja välineisiin, joilla niitä analysoidaan, vaan myös ajattelutavan ja tutkimuksen laskennalliseen käänteeseen. Big data synnyttää tietomuodon, joka muuttaa tiedon kohdetta, ja tavan, jolla ihmisverkostoja ja yhteisöjä ajatellaan. Siten big data voi luoda ”radikaalin muutoksen siinä, miten ajattelemme tutkimusta” (emt., 665).

Kaikkea big data ei kuitenkaan selvästikään yhteiskuntatutkimuksessa (tai muussakaan tutkimuksessa) tule muuttamaan. Se ei poista tarvetta teorioille. Numerot itsessään eivät kerro meille mitään; tutkijat ovat edelleen datan tulkitsijoita. Big data ei poista kontekstin ymmärtämisen ja substanssikohtaisen asiantuntemuksen tärkeyttä. Ajatus, että data voisi puhua puolestaan (vrt. Mayer-Schönberger & Cukier 2013, 6–11) ilman ihmisten sotkeutumista asiaan vääristymineen ja viitekehyksineen, on virheellinen; aineisto ei ole koskaan puhtaan luonnollista, vaan se luodaan ja konstituoidaan aktiivisesti erilaisten instrumenttien avulla. Huolimatta pyrkimyksestä kaikenkattavuuteen (jonka usein uskotaan virheellisesti johtavan automaattisesti hyvään edustavuuteen ja validiteettiin) big datakaan ei ole ikinä todellisuutta sellaisenaan. Kuten kaikki muutkin aineistot, vaikka se sisältäisikin ”kaikki tapaukset” ja olisi siten kokonaisaineisto, aineisto on aina tietystä näkökulmasta tehtävän valinnan ja rajauksen tulosta ja siten sekä näyte (*sample*) että representaatio todellisuudesta (Kitchin 2014). Luonnollisesti myöskään aineiston suuri koko ei ole tae aineiston laadusta tai siitä, että se olisi virheetön tai edustava (boyd & Crawford 2012).

Silti voi olla hyviä perusteita ajatella, että big data tulee merkitsemään suurta muutosta yhteiskuntatieteissä. Yhteiskuntatutkimuksen kannalta houkuttelevuus syntyy big datan avaamasta mahdollisuudesta päästä tutkimaan ”kokonaan uudenlaisia aineistomuotoja (sekä kvantitatiivisia että kvalitatiivisia), jotka ovat runsaita, käyttökelpoisia, nyansoituja, korkealaatuisia ja koottu mittasuhteissa, jotka ylittävät kapasiteettimme analysoida tai ymmärtää niitä”. Tämä voi merkitä

”historiallista muutosta yhteiskuntatieteille, joita on niiden alusta alkaen haitannut moninaiset datan niukkuuteen, ohuuteen ja kalleuteen liittyvät vaikeudet.” (Mohr ym. 2013, 676.)

Yksi keskeinen väärinkäsitys, joka on saattanut ohjata sosiologien kiinnostusta tai kiinnostumattomuutta big datasta on se, että big data merkitsisi murrosta tai edes haastetta ensi sijassa ”kvantitatiiviselle” yhteiskuntatieteelle. Esimerkiksi edustavaan otokseen perustuvat kyselytutkimukset tai rekisteriaineistoihin perustuvat tutkimukset eivät ole katoamassa mihinkään. Big data tulee niiden rinnalle; näköpiirissä olevassa tulevaisuudessa big data ei voi millään kattaa moniakaan niistä sisällöistä, joita näillä perinteisillä tavoilla yleensä tutkitaan. Suuri osa datatulvaa syntyy internetin ja sosiaalisen median sivutuotteena, eikä tuota dataa voida läheskään aina käyttää sosiaalitutkijoiden pohtiman reaalimaailman (esimerkiksi. terveys, elintavat) ongelmien tutkimiseen. Silloin kun tämä on mahdollista (esimerkiksi kulutusta voisi tutkia luottokorttiostotietojen perusteella), ei ole selvää, että sosiaalitutkijoilla tulisi olemaan pääsyä kyseiseen aineistoon. Näkemyksemme mukaan big datan syvällekäyvimmät muutokset voivatkin syntyä erilaisten tekstiaineistojen analysoimista harrastavalle tutkimukselle, jota on yleensä totuttu pitämään ”kvalitatiivisena”. Itse asiassa big data voi osaltaan murtaa kvalitatiivisen ja kvantitatiivisen tutkimuksen välisen jaottelun merkitystä; yhtäältä siksi, että se tuo laskennallisen, merkityksiä mittaavan analyysitavan perinteisesti ”kvalitatiivisiksi” miellettyjen tekstiaineistojen analyysirepertuaariin, toisaalta taas siksi, että näin tehdessään se luo suoran linkin, tai ainakin mahdollisuuden linkkiin, perinteisen kvalitatiivisen tulkintatavan ja formaalin mallinnuksen välille (esim. Light 2014; Mohr & Bogdanov 2013).

Toinen merkittävä ja toistaiseksi melko vähälle huomiolle jäänyt seikka on se, että big datan nousuun liittyvien, uusien laskennallisten tekstidatan analyysimenetelmien anti ja käyttömahdollisuudet eivät rajoitu vain poikkeuksellisten suurien, kirjaimellisesti ymmärrettyjen big datojen analyysiin. Ne avaavat – kuten tässä artikkelissa pyrimme argumentoimaan – lupaavia mahdollisuuksia ja merkitsevät potentiaalisesti syvällekäyvää muutosta myös ”tavallisen” kokoisten tai ”pienen datan” tekstiaineistoja analysoivan sosiologisen tutkimuksen kannalta, jota aiemmin on pidetty yleensä aina kvalitatiivisena tutkimuksena (vrt. Mohr & Bogdanov 2013, 561; Mohr ym. 2013).

Tekstiaineistojen analyysitavoista

Sosiologiassa käytetyt perustavat analysoida tekstidataa voidaan jaotella eri tavoin (ks. esim. DiMaggio, Nag & Blei 2013; Leetaru 2012; Light 2014). Perinteisessä mielessä on helppoa erotella toisistaan kaksi tarkastelutapaa (Biernacki 2014); yhtäällä on vapaa, aineistolähtöinen, yhden tutkijan tekemä tulkinta tai ”luenta” aineistosta, joka pyrkii mahdollisimman syvälliseen tulkintaan tai ”tiheään kuvaukseen” (Geertz 1973) aineiston merkityksistä. Tätä perinteistä kvalitatiivista (tai hermeneuttista) tulkintatapaa vasten asettuu usean tutkijan joukolla tekemä tekstiaineiston koodaus, joka perustuu ennalta sovittuihin luokittelu- ja koodausperiaatteisiin, joita pyritään systemaattisesti noudattamaan (Lasswell, Lerner & de Sola Pool 1952). Tämä jaottelu voidaan nähdä päällekkäisenä diskurssianalyysin ja sisällönanalyysin välisen erottelun kanssa (ks. Herrera & Braumoeller 2004), joista ensin mainittu (tai ainakin sen useimmat muodot) on paraatiesimerkki aineistolähtöisestä ”kvalitatiivisesta” tulkintaotteesta (Phillips & Hardy 2002) ja jälkimmäinen ”kvantitatiivisesta” sisällön erittelystä (Neuendorf 2002).

Näiden kahden perinteisen analyysitavan rinnalle ovat tulleet erilaiset tietokoneavusteiset ja enemmän tai vähemmän automatisoidut, koneelliset analyysitavat (Grimmer & Stewart 2013; Leetaru 2012). Olennaisena lähtökohtana niissä on, että

tekstiaineistoja analysoimalla voidaan *mitata* merkitysrakenteita (Mohr 1998; Mohr & Ghaziani 2014). Taustalla olevan metateoreettisen realistis-konstruktionistisen näkökulman mukaan merkityksiä ja diskursseja voidaan analysoida ja formalisoida aivan yhtä hyvin (tai huonosti) kuin mitä tahansa muitakin sosiaalisia ilmiöitä (ks. Ignatow 2015). Koneelliset analyysitavat voivat olla joko aineistolähtöisiä (ja siten induktiivisia) tai valmiiseen skeemaan perustuvia siinä missä ihmisen toteuttamatkin (vrt. Light 2014). Sen, miten uudet laskennalliset analyysitavat tavallaan rikkovat vanhat vastakkainasettelut, voi todeta myös viimeaikaisesta keskustelusta, jossa traditionaalista hermeneuttista ja humanistista tulkintaa puolustavat tutkijat ovat kritisoineet ”koodaamista” ja objektiivisuuteen pyrkiviä luokitteluita (Biernacki 2012; 2014). Kritiikin mukaan ennalta päätetyt koodit eivät ”löydä” aineistosta mitään sen omilla ehdoilla, vaan paremminkin pakottavat etukäteisskeeman mukaiset tulkinnat ulos aineistosta. Mutta jos analyysitavan formalisoinnin astetta *nostetaan vielä korkeammalle* ja käytetään automaattisia tekstidatan analyysimenetelmiä, jotka eivät sisällä tulkintaa (koodeja) ennen aineiston analyysia vaan etenevät induktiivisesti ja aineistolähtöisesti, esitetty kritiikki ei enää päde (Lee & Martin 2015). Nämäkin formaalit tekniikat tietenkin yksinkertaistavat tekstiaineistojen sisältöä, mutta tekevät sen samalla tavalla kuin kartta yksinkertaistaa maastoa; ne tekevät rakenteet näkyviksi ja mahdollistavat yhteiset tutkimusmatkat (emt.).

Kun koneelliset menetelmät otetaan huomioon, erottelu kvalitatiivisen ja kvantitatiivisen välillä ei ole siis enää keskeisin saati ainoa tapa jaotella tekstidatan analyysimenetelmiä. Nähdäksemme vaihtoehtoiset tekstidatan analyysitavat voidaan sen sijaan piirtää uudessa tilanteessa parhaiten nelikenttään, joka jakautuu toisaalta sen mukaan, onko analyysi (koodaus) tulosta ihmisen vai automatisoidun prosessin (koneen) työstä, ja toisaalta sen mukaan, perustuuko analyysi aineistolähtöisyyteen vai valmiin tulkintaskeeman noudattamiseen. Esitämme nelikentän kuviossa 1.

Ruudun 1 perinteisen ”kvalitatiivisen” luku- tai analyysitavan, jossa yksi ”virtuoositutkija” (vrt. DiMaggio, Nag & Blei 2013, 577) tulkitsee ja analysoi tiettyä tekstiä, etuna on mahdollisuus ”syvällisiin”, ”tiheisiin” ja spesifeihin tulkintoihin sekä prosessien ja narratiivien tavoittamiseen (ks. esim. Light 2014, 113). Heikkoudet ovat ilmeisiä: tulkintojen lähtökohdat ovat enemmän tai vähemmän arbitraarisia (samaa tekstiä voi lähestyä hyvin monelta eri analyysikulmalta, eikä valinnalle eri vaihtoehtojen väliltä ole helppoa antaa rationaalista perustetta), analyysi on viime kädessä subjektiivinen (ei ole selvää, että toinen tutkija tekisi samoistakaan lähtökohdista samankaltaisia tulkintoja; analyysillä ei siis ole toistettavuutta) eikä tällainen tulkitseva analyysi voi kattaa kuin hyvin rajallisen kokoisia tekstiaineistoja.

Ruudun 2 sisällönanalyysi perustuu useamman tutkijan (koodaajan) toteuttamaan, systemaattisuuteen pyrkivään, tiettyyn aiemman tutkimuksen ja teoreettisten perusteiden nojalla ennalta lukkoon lyötyyn koodijärjestelmään, jota noudattaen tekstiaineistoa koodataan (ja tuloksena syntyvää aineistoa analysoidaan jälkikäteen). Etuna on, että menetelmällä voi kattaa jo huomattavasti suurempia aineistomassoja kuin ruudun 1 analyysitavalla (muttei kuitenkaan todella isoja aineistokorpuksia), koodaajien välistä reliabiliteettia voidaan testata (mikä on välttämätöntä, koska pyritään objektiivisuuteen ja jopa toistettavuuteen) sekä se, että valmis aineisto mahdollistaa tilastolliset analyysit. Analyysitapa on kuitenkin kallis, hidas ja hankala. Heikkoutena voi pitää myös tulosten ja tulkintojen vähäistä ”syvällisyyttä”, koska objektiivisuuspyrkimys koodeja rakennettaessa johtaa siihen, että koodit eivät voi olla kovin monimutkaisia tai -tulkinnallisia; ”mitä kiinnostavampia tutkimuskysymykset ovat analyyttisesti, sitä vaikeampaa on saavuttaa tyydyttävää koodaajien välisen reliabiliteetin tasoa” (DiMaggio, Nag & Blei 2013, 577). Keskeistä analyysitavalle on, että siinä edelly-

		KOODAAJA: Ihminen	KOODAAJA: Kone
TULKINTA-PERIAATE: Aineistolähtöinen/vapaa		1. Perinteinen kvalitatiivinen tulkinta, ”tiheä kuvaus”	3. Ohjaamaton (*unsupervised*) induktiivinen automatisoitu tekstianalyysi (esim. aihemallianalyysi)
TULKINTA-PERIAATE: A priori -skeemaa noudattava		2. Koodijärjestelmää noudattava sisällönanalyysi	4. Ohjattu (*supervised*) automatisoitu tekstianalyysi (esim. sentimenttianalyysi)

KUVIO 1. Tekstianeistojen analyysitavat koodaajan ja tulkintaperiaatteen mukaan.

tetään, että tutkija tietää jo ennen varsinaista analyysiä, mitkä asiat ovat löytämisen arvoisia tekstistä (emt.). Sen jälkeen kun koodauksen kategoriat on suunniteltu, testattu ja kirjoitettu sääntökirjaan ja ryhmä aloittaa korpuksen koodaamisen, ei voi enää kääntyä takaisin; tämä tarkoittaa, että ennen koodauksen (laskemisen) aloittamista, tutkijoilla täytyy olla vahva tietämys tutkimuksensa kohteesta. (Mohr & Bogdanov 2013, 562.)

Näiden perustapojen rinnalle ovat syntyneet automatisoidut (koneelliset) tekstidatan analyysimenetelmät (ruudut 3 ja 4), joista aineistolähtöisesti ja ”induktiivisesti” etenevät menetelmät (kuten aihemallit) ovat tässä artikkelissa lähemmän tarkastelun kohteena. Neljännen ruudun analyysivaihtoehto viittaa sellaisiin osittain automatisoituihin analyysimenetelmiin, joissa ihmiskoodaaja aluksi ”opettaa” koneelle tietyn koodisysteemin, jota kone alkaa sitten toistaa. Esimerkki tällaisista ohjatuista (*supervised*) analyysimenetelmistä on sentimenttianalyysi. Tilanpuutteen vuoksi tämä analyysivaihtoehto sivuutetaan tässä yhteydessä ilman lähempää tarkastelua (ks. kuitenkin esim. Bail 2014; Leetaru 2012). Voidaan olla myös sitä mieltä, että sosiologian (ja etenkin kulttuurisosiologian) kannalta ei-ohjatut (*unsupervised*) automaattiset tekstidatan analyysimenetelmät ovat lupaavampia kuin nämä ohjatut menetelmät (DiMaggio, Nag & Blei 2013, 577).

Molempia koneistettuja analyysitapoja (ruudut 3 ja 4) luonnehtivat kuitenkin seuraavat kolme keskeistä etua (Leetaru 2012, 2–3). Ensimmäinen etu on helposti saavutettavissa oleva korkea reliabiliteetti; parhaitenkin koulutetuilla ihmiskoodaajatiimeillä, toisin kuin tietokoneella, työn jälki vaihtelee; tietokone koodaa ja luokittelee täsmälleen samoilla kriteereillä koko ajan, työn kestosta tai aineiston koosta riippumatta; koneella ei ole myöskään ”koodaajan sisäisiä” (*intra-coder*) ongelmia reliabiliteetin kanssa, joka vaivaa ihmiskoodaajia. Toisin sanoen yksittäiset koodaajatkaan eivät tee tasalaatuista jälkeä, sen ohella, että koodaajien välillä on eroja. (Emt., 77). Toinen etu on mahdollisuus toistaa analyysi. Vaikka ihmistiimillä olisi kuinka hyvä koodikirja, on selvää, että tulokset voivat vaihdella, jos työ tehtäisiin

uudestaan; samaa sääntöjoukkoa noudattava tietokonepohjainen koodaussysteemi tuottaa aina täsmälleen saman tuloksen. Kolmas etu on skaala ja laajuus; vaikka kyseessä olisi hyvin rahoitettu projekti, isokaan ihmiskoodaustiimi ei pysty käsittelemään todella suuria aineistoja; tietokoneilla toteutettu koodaus on nopeaa (tulokset tulevat lähes silmänräpäyksessä) ja niillä voi kattaa lähes rajattoman suuria tekstiaineistoja.

Ohjaamattomien automatisoitujen analyysitekniikoiden ominaispiirre ja etu on – paremman systemaattisuuden, reliabiliteetin, toistettavuuden ja analysoitavaksi soveltuvien aineistojen isomman skaalan lisäksi – se, etteivät ne tarvitse *a priori* -tyyppisiä koodijärjestelmiä, vaan analyysimetodi pelkistää tekstidatan piileviä rakenteita induktiivisesti, ilman että ihmisten etukäteen tekemiä luokitteluita tai koodauksia tarvittaisiin mallintamiseen (Bail 2014; ks. myös Blei 2012a; Mohr & Bogdanov 2013). Aihemallit (ja sen yleisimmin käytetty ja yksinkertaisin muoto, *Latent Dirichlet Allocation*, LDA; ks. Blei, Ng & Jordan 2003) tarjoavat automatisoidun menetelmän tekstikorpuksen koodaamiseksi joukoksi merkityksellisiä koodauskategorioita, joita kutsutaan ”aiheiksi”. LDA:n algoritmi tekee tämän ilman ihmisten interventiota (tai vain minimaalisella interventiolla), mikä tekee menetelmästä induktiivisemman kuin perinteiset tekstianalyysien menetelmät kulttuuri- ja ihmistieteissä. Sen sijaan, että lähtökohtana olisivat valmiit koodausskeemat, tutkija määrittää vain aiheiden lukumäärän, joka algoritmin tulee löytää. Tämän jälkeen ohjelma tunnistaa kyseisten aiheiden määrän ja tuottaa todennäköisyyslaskelman kullekin sanalle esiintyä kussakin aiheessa, sekä laskee aiheiden jakauman tekstikorpuksen sisällä (Mohr & Bogdanov 2013, 546).

Kuvion 1 avulla on hyvä pohtia eri tekstianalyysitapojen suhdetta toisiinsa nähden. Uusien koneellisten analyysimenetelmien, etenkin ohjaamattomien mallien, nousun merkityksen voi nähdä siinä, että sen ja perinteisen ”kvalitatiivisen” tekstiaineiston tulkinnan välillä on toisiaan täydentävä, komplementaarinen suhde (vrt. Blok & Pedersen 2014; Janasik, Honkela & Bruun 2009; Light 2014; Zamith & Lewis 2015). Kuten seuraavassa aihemallien ominaispiirteitä tarkemmin käsittelevässä jaksossa sekä aihemallianalyysiä demonstroivassa presidentin uudenvuodenpuheaineiston analyysissä esitämme, aihemallit eivät tule korvaamaan perinteistä ”tiheää kuvausta” ja tekstin lähiluentaa. Sen sijaan kuvion 1 ruudun 2 kvantitatiivinen sisällönanalyysi voi olla suhteellisen epäkäytännöllisenä ja työläänä tekniikkana tulevaisuudessa enemmän vaakalaudalla uusien koneellisten analyysitapojen kehittymisen vuoksi, koska automatisoidun analyysin selvinä etuina ovat analyysin ekonomisuus, luotettavuus ja mahdollisuus analysoida hyvin suuria aineistoja (ks. kuitenkin Lewis, Zamith & Hermida 2013; Zamith & Lewis 2015). Palaamme tekstianalyysitapojen väliseen työnjakoon artikkelin lopussa.

Aihemallit esimerkkinä uusista laskennallisista tekstidatan analyysimenetelmistä

Mitä aihemallit ovat?

Aihemallit *(topic models*; Blei, Ng & Jordan 2003; Blei 2012a) olettavat, että yksittäinen dokumentti koostuu useammasta aiheesta, joita vuorostaan määrittävät sanat, tai tarkemmin sanojen todennäköisyydet. Aineiston eli havaittujen dokumenttien perusteella mallinnetaan samanaikaisesti kaksi asiaa: aiheiden jakauma dokumenteissa ja sanojen jakauma aiheissa. Näiden voi ajatella kuvaavan tekstit generoivaa prosessia: jos dokumentin kirjoittaja rakentaisi tekstinsä siten, että ensin valittaisiin aiheet ja sen jälkeen satunnaisesti poimittaisiin aiheisiin liittyviä sanoja jakaumista, nämä jakaumat ovat ne, jotka todennäköisimmin synnyttäisivät nyt havaitut dokumentit. Aihemallit ovat siis niin sanottuja *mixed membership* -malleja, joissa jokainen dokumentti

koostuu useammasta aiheesta eikä tavoitteena ole erotella dokumenttityyppejä.

Mallin ensimmäisenä tuotoksena on jokaisen aiheen todennäköisyysjakauma kaikkien aineistossa esiintyvien sanojen yli, eli millä todennäköisyydellä kyseisestä aiheesta sanaa poimittaessa tulisi juuri tämä sana. Jokaiselle dokumenteissa esiintyvälle sanalle on laskettu todennäköisyys kussakin aiheessa – sanat voivat siis ilmetä useammassa eri aiheessa, mikä tarkoittaa, että aihemallit tunnistavat sanojen monimerkityksisyyden. Mallin tulkinnassa tarkastellaan tätä jakaumaa tai yleensä siinä suurimman todennäköisyyden saaneita sanoja; onko esimerkiksi kymmenelle yleisimmälle sanalle aiheessa jokin merkityksellinen tulkinta ja mitä tämä kertoo tekstiaineiston rakenteesta. Tämä vaihe on analoginen faktorianalyysin tulkinnan kanssa: aiheet vähintään nimetään jatkoanalyyseja varten ja mahdollisesti niiden ominaisuuksiin pureudutaan syvällisemminkin. Toinen tuotos on kunkin dokumentin jakautuminen aiheisiin eli kuinka iso osa dokumentin sanastosta on mistäkin aiheesta.

Mallin generoiva prosessi ei tietenkään kuvaa tekstien todellista syntyprosessia, mutta estimoidut jakaumat voivat kertoa jotain mielenkiintoista tekstikorpuksen merkitysrakenteen perusluonteesta (Mohr & Bogdanov 2013; Mützel 2015). Ainakin ne voivat toimia hakualgoritmeina: jos vaikkapa haluaa lukea lehden urheilusivuilla pesäpalloa käsitteleviä artikkeleja, aihemalli saattaisi kertoa, että yhdessä aiheessa korkean todennäköisyyden saavat ”lukkari”, ”palo” ja ”sisävuoro”, ja lukija voisi poimia ne artikkelit, joissa iso osa sanoista on tästä aiheesta. Yhteiskuntatieteellinen tutkimus aihemalleilla vaatii lisäksi tulkintateoriaa, joka kertoo, mitä yhteiskunnallista ilmiötä tai mekanismia aiheet kuvaavat. Kirjallisuudessa aiheiden on tulkittu olevan muun muassa kehysanalyysin kehyksiä (DiMaggio, Nag & Blei 2013), politiikan dramatiikan näyttämöjä (Mohr ym. 2013) tai poliittisia agendoja (Grimmer 2010).

Sosiologit ovat löytäneet aihemallit analyysivälineikseen vasta aivan viime vuosina, mutta voidaan odottaa, että aihemalleja käyttävät tutkimukset tulevat lisääntymään. Aihemalleja on hyödynnetty toistaiseksi ennen muuta kulttuurisosiologiassa mutta myös esimerkiksi taloussosiologiassa sekä yhteiskunnallisten liikkeiden ja politiikan tutkimuksessa. Kulttuurisosiologinen esimerkki on DiMaggion ja kumppaneiden (2013) kattavaan sanomalehtiaineistoon perustuva analyysi taiteen julkisen rahoituksen politisoitumisesta Yhdysvalloissa 1980- ja 1990-luvuilla. Taloussosiologi Neil Fligstein kumppaneineen (2014) on ottanut aihemallianalyysin kohteeksi puolestaan Yhdysvaltain keskuspankin ohjauskorosta päättävän avomarkkinakomitean vuosien 2000–08 välisten kokousten sanatarkat keskustelupöytäkirjat, joiden analysoiminen valottaa, miksi vuoden 2008 finanssikriisi yllätti rahoitusmaailman asiantuntijat ja mistä syystä säätelevät toimenpiteet olivat niin hitaita. Aihemallianalyysin kohteena ovat olleet siis hyvin monenlaiset tekstiaineistot: niin sanomalehti- kuin sosiaalisen median artikkelit, erilaiset asiakirjat ja muut julkaisut sekä nauhoitetut keskustelut ja puheet.

Keskustelu aihemallien rajoista ja mahdollisuuksista yhteiskuntatieteissä on vasta alussa. Rajoituksien osalta on tärkeää ymmärtää mallin takana olevat oletukset, jotka liittyvät aineiston esikäsittelyyn. Aihemallit perustuvat (ainakin perusversiossa, LDA:ssa) sanojen esiintymiseen samassa dokumentissa, riippumatta siitä missä kohtaa tai minkälaisessa yhteydessä ne esiintyvät. Sanajärjestys ja kaikki kielioppi jätetään siis huomiotta – dokumentit ovat sanasäkkejä (*bag-of-words*), jotka on helppo kuvata dokumentissa esiintyvien sanojen frekvenssijakaumina. Sanojen merkitys malleissa syntyy siis täysin siitä, minkä muiden sanojen kanssa niitä käytetään. Oletus on epärealistinen, vaikeuttaa joskus mallien tulkintaa ja tarkoittaa, että aihemallit eivät ole varsinaisesti malleja kielenkäytöstä. Kuinka vakavia ongelmia tämä aiheuttaa, riippuu tutkimuskysymyksestä ja

tavoitteista. Jos tavoitteena on ennakoiva tekstinsyöttö hakukoneessa, sanajärjestys on tietysti keskeinen, mutta jos tavoitteena on kuvailla, mitä aiheita kukin miljoonasta tieteellisestä artikkelista käsittelee, kieliopilla ei ole juurikaan merkitystä.

Toinen mallin oletus tai rajoite ovat sen vaatimukset tekstin esikäsittelylle ja niiden vaikutus tuotoksiin. Raakatekstistä poistetaan hyvin tavalliset sanat (suomen kielessä poistettavia sanoja ovat siis esimerkiksi ”ja”, ”tai”, ”edes”, ja niin edelleen) ja sanat stemmataan eli katkaistaan juurimuotoonsa tai lemmatisoidaan eli muutetaan perusmuotoonsa. Lemmatisointi vaatii täydellisen sanakirjan taivutusmuotoineen, ja kevyempi stemmaus on siksi usein mielekäs (Järvelin & Pirkola 2005). Toimenpiteet ovat kielikohtaisia ja joskus melko vaativia – suomen kielen luonne on sellainen, että mielenkiintoisen informaation säilyttämiseksi ei ole aina selvää, kuinka paljon sanoja kannattaa muuttaa perusmuotoon. Lisäksi tekstistä poistetaan välimerkit, mahdollisesti numerot, isot kirjaimet ynnä muuta sellaista. Tämä operaatio ei riipu kielestä ja on verraten yksinkertainen. Joskus on tarpeen poistaa myös liian yleisiä tai liian harvinaisia sanoja, erityisesti jos aineisto on verrattain pieni. Esimerkiksi muutamassa dokumentissa runsaasti käytetyt sanat saattavat saada mallin luulemaan näiden olevan keskeisiä aiheita. Raakakäsittelyn jälkeen dokumentit muutetaan dokumentti-käsite-matriiseiksi, taulukoiksi, joissa listataan, kuinka monta kertaa kukin analyysiin mukaan haluttu sana esiintyy kussakin dokumentissa.

Esikäsittelyn jälkeen itse mallinnuksen suhteen on tehtävä kaksi päätöstä: ensiksi aihemallinnuksessa tutkijan täytyy valita haluamansa aiheiden määrä – erilaisten valintojen vertailuun ei ole selkeitä tilastollisia työkaluja, vaan ainoastaan tulkinnan heuristiikat. Aiheiden määrä ei riipu aineiston koosta (vaikka luultavaa ehkä onkin, että hyvin suurissa aineistoissa on usein enemmän aiheita kuin pienissä), vaan aineiston fokusoituneisuudesta. Tutkijan täytyy päättää – ja viime kädessä kokeilla –, onko sopiva aiheiden määrä esimerkiksi 3, 20 vai 200. Lisäksi mallissa on hyperparametrit, jotka painottavat sitä, kuinka monesta aiheesta dokumentissa voi olla sanoja. Mallissa on myös satunnainen alkupiste. Nämä vaikuttavat tuloksiin, ja aihemallinnuksen kaltaisiin menetelmiin kuuluukin luonnostaan iteratiivinen rakentelu, tulkinta, muutokset ja parantelu (Blei 2014). Myös erilaisten mallien kokeilu on mielekästä.

Aihemallien suosion nopea kasvu johtuu osittain niiden laajennettavuudesta: perusmallin päälle ja ympärille on mahdollista rakentaa monipuolisesti realistisempia ja erilaisten aineistojen ominaisuuksia kunnioittavia malleja. Näistä esimerkkejä ovat aikasarja-aihemallit, joissa sanastot aiheiden sisällä voivat kehittyä ajan myötä (Blei & Lafferty 2006), erilaiset sanasäkkioletuksen kevennykset ja laajennukset sanoista fraaseihin (esim. Danilevsky ym. 2013) sekä mallien ennustuskyvyn arvioinnin ja hyödyntämisen tekniikat (Chang ym. 2009).

Esimerkki aihemallinnuksesta: tasavallan presidenttien uudenvuodenpuheet

Konkreettinen esimerkkimme aihemallisovelluksesta on tarkoituksellisesti yksinkertainen: sen tuli olla toteutettavissa verraten yksinkertaisin työkaluin, pyrkimyksenä oli valita aihepiiri ja aineisto, jota koskevan analyysin tulokset olisivat suhteellisen helposti avautuvia ja intuitiivisia suomalaiselle sosiologiyleisölle mahdollisimman vähäisellä johdattelulla aihealueen taustoihin ja historiallisen kontekstin yksityiskohtiin. Suomen tasavallan presidenttien uudenvuodenpuheista koostettu aineisto täyttää nämä ehdot. Tämän artikkelin tila ei riittäisi, jos käyttäisimme aineistoa, joka pitäisi kontekstualisoida alusta alkaen. ”Oikeissa” aihemalleja käyttävissä tutkimuksissa analyysin kontekstualisoinnin ja tarkasteltavaa kenttää koskevan taustatietämyksen merkitystä ei voi aliarvioida (DiMaggio, Nag & Blei 2013; Mohr 1998). Vakavampi analyysi presidentinpuheistakin edellyttäisi kontekstin syventämistä tuomalla esiin kutakin puheenpitohetkeä koskevaa

lisätietoa tulkinnan terävöittämiseksi. Mohrin (1998, 366) mukaan ”paras peukalosääntö” onkin ”paikantaa ja arvioida relevanttia käytännön toiminnan aluetta, johon tunnistetut kulttuuristen merkitysten systeemit ovat uponneet”.

Suomen tasavallan presidentit ovat pitäneet radioidun tai televisioidun uudenvuodenpuheen vuodesta 1935 alkaen, jolloin presidentti Svinhufvud aloitti perinteen. Kotimaisten kielten tutkimuskeskus on kerännyt puheet vuoteen 2006 asti ja jakaa korpusta keskuksen Kaino-palvelussa. Aineistomme koostuu tästä korpuksesta sekä lisäksi kahden viimeisimmän presidentin kotisivuillaan jakamista puhekäsikirjoituksista ja kattaa siten kaikki 81 uudenvuodenpuhetta (myös ne, jotka on pitänyt ministeri tai pääministeri presidentin sijasta). Aineiston peruspiirteet on tiivistetty taulukkoon 1.

Presidenttien uudenvuodenpuheita on tutkittu myös aiemmin (Heikkinen 2006; Heikkinen & Lounela 2008; vrt. myös Light 2014, jonka analysoimat Yhdysvaltain presidenttien virkaanastujaispuheet ovat eräänlainen vastinpari tälle kotimaiselle aineistolle). Poliittisen kielen muuttumista koskevassa analyysissaan kielitieteilijät Heikkinen ja Lounela (2008) analysoivat korpusta kvantitatiivisten morfologisten menetelmien avulla sekä kvalitatiivisesti tekstien sisältöä tarkastellen. Tutkijalähtöisesti he myös koodaavat tekstin aiheet, jotka heidän analyysissaan ovat kotimaa, maailma ja yleiskategoria, ja vertailevat esimerkiksi erilaisten verbien yleisyyttä ja ensimmäisessä persoonassa puhumista näitä aiheita käsittelevissä teksteissä. Heikkisen ja Lounelan tutkimuksessa aihe on tekstijaksoa koskeva mittari, kun taas tässä tehtävä aihemallinnus koskee yksittäisiä sanoja ilman sanajärjestystä. Yhteiskuntatieteellisessä analyysissamme mielenkiinnon kohteena ei ole kielenkäyttö itsessään, vaan puheiden kuvaamat poliittisesti ja yhteiskunnallisesti tärkeät asiat: tulla mainituksi tai käsitellyksi

TAULUKKO 1. Tasavallan presidenttien uudenvuodenpuheaineiston peruspiirteet

Presidentti	Vuodet	Puheita	Puheen pituus (sanaa, ka)
Svinhufvud	1935–1937	3	277
Kallio	1938–1940	3	1336
Ryti a	1941–1944	4	1024
Mannerheim b	1945–1946	2	758
Paasikivi	1947–1956	10	918
Kekkonen	1957–1981	25	907
Koivisto c	1982–1994	13	1077
Ahtisaari	1995–2000	6	1014
Halonen	2001–2012	12	1015
Niinistö	2013–2015	3	837
Yhteensä	1935–2015	81	77600 (koko aineisto)

a Vuonna 1942 puheen piti Rydin asemasta eduskunnan puhemies Väinö Hakkila ja vuonna 1944 pääministeri Edwin Linkomies.

b Mannerheim ei pitänyt uudenvuodenpuheita; vuonna 1945 puheen piti kansliaministeri Mauno Pekkala ja vuonna 1946 tuolloin vt. presidenttinä toiminut pääministeri Paasikivi.

c Ensimmäisen puheensa vuonna 1982 Koivisto piti vt. presidenttinä toimineena pääministerinä. Vuonna 1993 puheen piti Koiviston asemasta pääministeri Esko Aho.

presidentin uudenvuodenpuheessa on jo itsessään osoitus asian poliittisesta merkittävyydestä.

Tässä esiteltävä analyysi ja siihen liittyvä esikäsittely on toteutettu R-kielellä ja tm- ja topic models -paketeilla, joista ensimmäinen tarjoaa työkaluja tekstitiedostojen käsittelyyn monilla kielillä, myös suomeksi, ja jälkimmäinen varsinaiset aihemallien sovellukset. Aineisto stemmattiin Snowball-stemmausalgoritmilla ja sen R-implentaatiolla SnowballC-paketissa. Lisäksi karsittiin hyvin tavalliset sanat, typografiset elementit sekä muutamia uudenvuodenpuheen rakenteeseen liittyviä käsitteitä ja sanoja, jotka esiintyvät erittäin harvoin tai (melkein) kaikissa puheissa.

Lopputuloksena on matriisi, jossa on 81 puhetta ja 5 557 analyysiin valikoitunutta sanaa sekä tieto siitä, kuinka monta kertaa kussakin puheessa jokaista sanaa käytetään. Tähän matriisin sovitettiin aihemalleja varioiden mallinnettujen aiheiden määrää sekä mallin hyperparametrejä. Mallit ovat tulkinnallisesti varsin samanlaisia, kunhan aiheita on riittävästi mutta ei liikaa. Muutaman aiheen mallit erottelivat aineistosta aikakausia, laajemmat yksittäisten presidenttien puheiden aiheita ja tyylejä. Seitsemän aiheen mallissa yhdistyy sekä aikakausien tärkeitä teemoja että historian kaikkina aikoina puhuttavia asioita. Malli valittiin tässä esitettäväksi, koska mielenkiinnon kohteena on pohtia historian kaarta, ei esimerkiksi ennustaa tekstin perusteella kenen puheesta on kyse. Yhden aiheen lisäys tai vähennys ei kuitenkaan merkittävästi muuttaisi tulkinnan pääasioita, ja tarkka määrä on tutkijan heuristinen päätös. Mallin tulkinta ei muuttunut hyperparametrejä muuttamalla.

Esimerkkianalyysin tulokset: tasavallan presidentin uudenvuodenpuheet 1935–2015

Taulukko 2 esittää kunkin aiheen kymmenen yleisintä käsitettä siten, että muutamasta aiheesta on poistettu saman sanan eri muotoja, joita käytetty stemmausalgoritmi ei tunnistanut. Liitetaulukossa[1] esitetään puolestaan aiheiden jakautuminen puheisiin, eli se osuus kunkin puheen analyysiin mukaan luetuista sanoista, joka on kustakin aiheesta. Seitsemän aiheen malli kertoo, miten uudenvuoden puheet ovat institutionalisoituneet, mitä ne ovat käsitelleet ja miten niiden merkitys on rakentunut. Taulukon 2 sarakkeiden otsikot ovat tekemiämme tulkintoja. Tulkinnoissa on käytetty 50 yleisintä käsitettä kustakin aiheesta.

Ensimmäisen aiheen sanasto on yleistä kansallistunnon nostoa ja juhlapuhekieltä. Erityisesti aineiston ensimmäiset, Svinhufvudin ja Kallion puheet käsittelevät lähinnä tätä aihetta. Uudenvuodenpuhe ei ollut ehkä vielä saanut sitä yhteiskunnallista merkitystä, joka sille rakentui myöhemmin, ja puheet ovat aidosti vain presidentin tervehdyksiä kansakunnalle. Juhlapuhesanastoa käytetään myöhemminkin, mutta se muodostaa puheitten pidetessä pienemmän osan kokonaissanastosta (vrt. Taulukko 1). Esimerkiksi käy aivan ensimmäinen uudenvuodenpuhe, jonka piti Svinhufvud vuonna 1935 (tässä ja seuraavissa lainauksissa sanoihin merkityt yläindeksinumerot vastaavat aiheita, joihin malli on kyseisen sanan sijoittanut): ”Täten ulkonaisestikin osoitamme lujan[1] päätöksemme yhteisvoimin[1] ponnistella[4] vaikeuksia vastaan, jotka edelleen järkyttävät maailmaa, uhaten[5] turmiolla[1] heikkoja[2] ja epäröiviä[1].” Mallin ja tehtyjen valintojen ominaisuudet näkyvät hyvin lainauksessa: suurin osa sanastosta on mallinnettu aiheeseen, jota lause intuitiivisesti käsittelee, mutta osa sanoista myös toisiin: heikkoudesta puhutaan muualla aineistossa erityisesti talouskehityksen yhteydessä – jos heikoista olisi puhuttu useammin myös tässä merkityksessä, malli olisi voinut eritellä merkitykset omiin aiheisiinsa. Jotkut merkitykselliset sanat jäävät myös mallittamatta: maa-

1 Liitetaulukko on julkaistu sähköisesti osoitteessa http://sosiologia.fi/pdf/purhonen-toikka2016.pdf

TAULUKKO 2. Uudenvuodenpuheaineiston aiheet sekä aiheiden kymmenen yleisintä sanaa ja sanojen poiminnan todennäköisyydet

1. Juhlapuhe	2. Talouspuhe	3. Talous-kriisipuhe	4. Vaalit ja puolueet	5. Ihmiset ja sosiaaliset kysymykset	6. Kriisit	7. EU+
luoto	milj	järkev	puolue	asunto	energia	union
(0,007)	(0,009)	(0,007)	(0,015)	(0,008)	(0,01)	(0,046)
tahdo	miljard	sisäpoliittis	vaal	kohtuullis	ydinas	laps
(0,006)	(0,009)	(0,005)	(0,01)	(0,008)	(0,01)	(0,016)
lausu	jälkeis	jäsenyyd	sotakorvauks	optimistis	kirj	itämer
(0,005)	(0,008)	(0,004)	(0,005)	(0,005)	(0,008)	(0,009)
nykyä	kustannus-taso	devalvaatio	johtan	juhlavuod	öljy	onnetto-muud
(0,005)	(0,005)	(0,004)	(0,005)	(0,004)	(0,008)	(0,005)
yhteisvoim	lisäys	kiina	menestyk-sellis	luotae	pakolais	perustuslak
(0,005)	(0,005)	(0,003)	(0,005)	(0,004)	(0,006)	(0,005)
luja	valtiovierailu	laki	päämäär	rehellis	palkansaaj	euro
(0,005)	(0,004)	(0,003)	(0,004)	(0,004)	(0,006)	(0,004)
miehe	kauppatas	maaseudu	vaale	kansantulo	gorbatshov	globaal
(0,005)	(0,004)	(0,003)	(0,004)	(0,003)	(0,005)	(0,004)
taistelu	korkea-suhdant	pystyy	erimielisyyt	tasaarvois	heiko	väkiv
(0,005)	(0,004)	(0,003)	(0,004)	(0,003)	(0,005)	(0,004)
ketä	odotuks	rakenne-muutoks	melkois	jäsenyysneu-vottelu	maapalo	kulutustaso
(0,004)	(0,004)	(0,003)	(0,004)	(0,003)	(0,005)	(0,004)
lukumäär	teollisuus-tuotan	reaaliansio	tapahtui	kehittym	ohjelm	terrorism
(0,004)	(0,004)	(0,003)	(0,004)	(0,003)	(0,005)	(0,004)

ilma on niin yleinen sana puheissa, että se putoaa pois sanoja karsittaessa; ulkonaisestikin taas on niin erikoinen muoto, että käytetyt työkalut eivät löydä sille mielekästä perusmuotoa, jolloin sana jää ainoaksi korpuksessa ja jää siten mallittamatta.

Kaksi seuraava aihetta koostuvat taloutta ja talouden tilaa koskevasta sanastosta. Erityisesti Kekkosella oli tapana avata puheensa tilastokatsauksella, jossa käytiin läpi valtiotalouden miljoonia, miljardeja, prosentteja ja kauppataseita. Toinen aihe rakentuukin voimakkaasti Kekkosen käyttämän sanaston kautta, vaikka aihetta käsitellään myös aikaisemmissa ja jossain määrin uudemmissa puheissa viimeisiä vuosikymmeniä lukuun ottamatta. Esimerkiksi 1957 talouden ti-

laa kuvattiin näin: ”elinkustannusindeksin[2] lähes 18 %:n nousu viime vuoden lopusta marraskuuhun[5], kauppataseen[2] 25 miljardiin[2] markkaan nouseva alijäämäisyys[2], valtion rahavaikeudet[1] ja suuri työttömien[2] määrä”.

Kolmas aihe koostuu hajanaisemmasta sanastosta, jonka yhdistävänä tekijänä on niiden liittyminen aikansa talouskriiseihin – rakennemuutoksesta, reaaliansioista, markkinatalousmekanismista ja lamatilanteesta puhutaan kautta aineiston, mutta aiheessa näkyy näiden erilainen merkitys eri aikoina: joskus aiheeseen yhdistyy puhe devalvaatiosta, joskus Kiinasta, joskus jäsenyydestä. Esimerkiksi Kekkonen veti kehitystä yhteen 1962: ”viime vuoden lopulla on eräiden vientituotteittemme[3] kysynnässä[3] ilmennyt hiljentymistä[3] ja hinnoissa[3] joitakin laskuja”.

Neljäs presidenttejä puhuttanut aihe liittyy vaaleihin ja puolueisiin sekä niihin liittyviin erimielisyyksiin ja pyrkimyksiin. Presidentit nostavat vaaliteemaa esille eri tavoin, kiitellen mandaatistaan, kehottaen äänestäjiä uurnille tulevana vuonna, tai jopa ottaen kantaa vaalitapaan – kuten Paasikivi puhuessaan vuonna 1951 hallituksen muodostamisen hankaluudesta: ”Komiteassa[4], joka vuoden vaihtuessa[2] 1905–1906 valmisti[2] ehdotuksen uudeksi valtiopäiväjärjestykseksi[4] ja vaalilaiksi[4], ja jonka jäsenistä minä olen ainoa elossa oleva, harkittiin vaalijärjestelmän[4] yhteydessä näitä asioita monelta puolelta”.

Viidennessä aiheessa presidenttejä puhuttavat sosiaaliset asiat ja suomalainen kansanluonne. Presidentit kuvaavat suomalaiset rehellisiksi ja luotettaviksi ja pohtivat myös sosiaalisia ongelmia, joista päällimmäiseksi tässä nousevat erilaiset asuntokysymykset asunnottomuudesta ja asuntojen hinnoista asuntolainojen alhaisiin korkoihin. Viides aihe on tullut käsitellyksi melko tasaisesti historian aikana, mutta nykyinen presidentti Niinistö on korostanut sitä erityisesti ja systemaattisemmin kuin esimerkiksi edeltäjänsä Halonen. Esimerkiksi vuonna 2014 Niinistö totesi: ”Meidän suomalaistenkin on oltava rehellisiä[5]. Minäkin olen; eurooppalainen[5] tie on meidän tiemme[1].”

Kuudes aihe sisältää sanastoa, jolla erityisesti Kekkonen ja Koivisto kuvailivat aikansa kriisitilanteita. Kylmä sota ydinaseineen sekä asevarustelusopimuksineen ja energiakriisi olivat aikakauden keskeisiä kysymyksiä. Koivisto kuvaili vuonna 1986 tyytyväisyyttään keskusteluun ”Pohjolan vakiintuneen ydinaseettomuuden[6] sopimuspohjaisesta[6] vahvistamisesta”. Seitsemäs aihe kerää yhteen viime vuosikymmenten teemoja ja yhdistää etenkin Ahtisaaren ja Halosen puheita: Euroopan unioni, globaalit kysymykset, terrorismi ja ympäristökysymyksistä erityisesti Halosen useammassakin puheessa esille nostamat Itämeren kysymykset. Nämä Eurooppa- ja globalisaatioaiheiset kysymykset ovat olleet puheiden keskeistä sisältöä Suomen EU-jäsenyydestä alkaen – sekä Lasten vuonna 1979 Kekkosen kehitysmaiden lasten asemaa käsitelleessä puheessa.

Tällä tavoin tarkasteltuna aihemallinnuksen avulla piirtyy kuva neljästä osittain päällekkäisestä aikakaudesta, joiden kanssa kolme teemaa kulkevat limittäin läpi aineiston. Dynaaminen malli, joka ottaisi ajan mukaan mallinnukseen, voisi sopia aineistoon vielä paremmin ja ideaalitapauksessa tunnistaa teemoja, joiden sisällä sanasto on muuttunut. Tässä esitetty yksinkertainenkin malli kertoo kuitenkin mielenkiintoisen tarinan puheista, jota olisi vaikea tehdä näkyväksi ainakaan etukäteen määritellyn koodiskeeman avulla.

Tässä käytetyn aineiston pieni koko ja kohtalaisen karkea sanaston esikäsittely asettavat kuitenkin rajansa mallin käytettävyydelle. Yleisimmissä sanoissa esiintyy jonkin verran teemoihin liittymättömiä sanoja, jotka saattavat nousta esiin vain

yhden puheen tai yhden presidentin tyylin perusteella. Tärkeämpää on kuitenkin se, että ratkaisu on kohtalaisen stabiili, eli sen perusrakenne ei vaihtele satunnaista alkupistettä tai painotusparametriä muokkaamalla. Yhteiskuntatieteilijöillä on paljon laadullisia aineistoja, jotka ovat suurin piirtein samaa kokoluokkaa kuin nyt käytetty puheaineisto. Tässä esitellyn esimerkin perusteella voidaan väittää, että aihemalleja on mahdollista käyttää vähintään tukemassa laadullista analyysia myös silloin, kun aineisto on näinkin rajallinen (Mohr ym. 2013; ks. kuitenkin Tang ym. 2014).

Aihemallit ja tekstiaineistojen analyysitapojen työnjako

On selvää, että aihemallien kaltaiset menetelmät ovat parhaimmillaan huomattavasti yllä käytettyä suurempien aineistojen analysoimisessa. Erilaisten tekstiaineistojen analyysitapojen (vrt. kuvio 1 edellä) työnjaon ja aineiston koon välillä onkin yhteys; hyvin suuret aineistot ovat analysoitavissa kokonaisuudessaan vain koneellisesti, ellei rajauduta vain pieniin otoksiin tai näytteisiin. Aihemallien kaltaiset uudet laskennalliset tekniikat siis laventavat yhteiskuntatieteilijöiden ja humanistien potentiaalista aineistorepertuaaria tuodessaan sellaiset tekstidatat mahdollisten analysoitavien aineistojen piiriin, joita ei niiden kokonsa vuoksi ole aiemmin voitu ajatellakaan tutkittavan kokonaisvaltaisesti (vrt. Bail 2014, 467–477). Pienempien aineistojen kohdalla tilanne on toinen. Toisessa ääripäässä, hyvin pienten aineistojen kohdalla, ei ole tietenkään mielekästä (tai edes teknisesti mahdollista) käyttää koneellisia analyysimenetelmiä, vaan pelkkä aineiston lukeminen ja tulkitseminen tutkijan toimesta riittää ja on luultavasti myös tehokkain analyysin tapa. Mutta kuten tässä käytetyn uudenvuodenpuheaineiston esimerkki osoittaa, aihemallit ovat käyttökelpoisia analyysin ja tulkinnan apuvälineitä jo melko rajallisten, yhteiskuntatieteissä hyvin tyypillisesti kerättävien ”kvalitatiivisten” tekstiaineistojen kohdalla, olivatpa kyseessä haastattelulitteraatiot, elämäntarinat, media- tai asiakirja-aineistot (vrt. Mohr & Bogdanov 2013, 561). Tällöin koneellisen ja perinteisen kvalitatiivisen analyysin suhdetta voi pitää toisiaan täydentävänä (vrt. Janasik, Honkela & Bruun 2009; Light 2014). Jos hyvin suurten aineistojen kohdalla komplementaarisuuden voi ajatella menevän niin, että aihemallianalyysin kaltaisella tekniikalla saadaan tiivistettyä perustietoa aineiston rakenteista (aiheista), jota voidaan sitten täydentää täsmentävillä ja ”tiheämmillä” tulkinnoilla valituista näytteistä, pienempien (tai sosiologiassa ”tavallisen” kokoisten) aineistojen kohdalla aihemallien voi ajatella olevan perinteistä tulkitsevaa analyysitapaa täydentävän ja tukevan tiedon lähde.

Aineiston koosta riippumatta (aivan pieniä aineistoja lukuun ottamatta) aihemallianalyysin kaltaisten ohjaamattomien koneellisten tekstiaineistojen analyysimenetelmien (kuvio 1, ruutu 3) komplementaarinen hyödyllisyys perinteiseen kvalitatiiviseen tulkintatapaan (kuvio 1, ruutu 1) verrattuna voidaan tiivistää kolmen seikan avulla. Ensinnäkin niiden avulla voi systematisoida aineistosta induktiivisesti nousevia, aineiston merkitysrakennetta aiheiden avulla kuvaavia havaintoja (yllä olevassa presidentinpuhe-esimerkissä mallinnuksen erittelemät seitsemän aihetta). Toiseksi aihemallit antavat mahdollisuuden systematisoida havaintoja aineiston merkityksistä aineiston sisäisten osajoukkojen mukaan (yllä esimerkiksi se, miten tunnistetut aiheet jakautuvat presidenttikausien mukaan, mutta yhtä hyvin ositus voisi tapahtua vaikkapa presidentin puoluekannan mukaan tai sen mukaan eletäänkö taloudellista nousu- vai laskukautta; jakoperuste voi olla mikä tahansa kyseisessä tutkimuksessa kiinnostuksen kohteena oleva, järkevästi jakautuva aineiston ominaispiirre). Kolmanneksi, eräänlaisena erityistapauksena aineiston osittamisesta, aihemallien kaltainen analyysi on parhaimmillaan silloin, kun käytössä on pitkittäisaineisto; aihemallien avulla voi systematisoida aineiston merkitysten muutoksen analyysiä.

Jokaisen kolmen seikan kohdalla aihemallit tarjoavat eräänlaisen selkänojan, johon nähden (perinteinen kvalitatiivinen) tutkija voi nojata ja tarkistaa omia tulkintojaan, jotka voisivat muutoin helposti jäädä liian kaavamaisiksi tai yksinkertaistaviksi. Aihemallit antavat tilaa myös yllätyksille ja detaljeille, jotka voivat jäädä ihmistulkitsijalta huomaamatta. Esimerkiksi ajallista muutosta koskevassa analyysissa aihemallianalyysi voi paljastaa, että viime aikoina yleinen aihe onkin esiintynyt myös jo aiemmin (vrt. Blei 2012a; DiMaggio, Nag & Blei 2013). Tällöin tutkija voi tarkistaa ja tulkita lähemmin, onko kyse todella samasta aiheesta samassa merkityksessä eri aikoina ja eri konteksteissa. Näin aihemallien käyttö perinteisen kvalitatiivisen tulkitsevan tutkimuksen rinnalla ja analyysin systemaattisuutta lisäävänä resurssina voi parhaimmillaan tarkoittaa täsmällisempiä ja luotettavampia tulkintoja. Jatkossa aihemallianalyysin kaltaiset laskennalliset tekniikat voivat tehdä etenkin sellaisen kvalitatiivisen tutkimuksen entistä vaikeammin perusteltavaksi, jossa esitetään vaikutelmanvaraiseksi jääviä, pohjimmiltaan laskemiseen perustuvia, aineiston (yleensä melko pienen) sisäisiä yleistyksiä aineiston piirteiden osaryhmittäisestä jakautumisesta (vrt. Töttö 2012). Aihemallien käyttö *ei* sen sijaan tietenkään vähennä saati poista tutkijan tekemien sisällöllisten ja mielekkäiden tulkintojen tarvetta (Blei 2012b).

Yllä on suhteutettu aihemallien antia nimenomaan perinteiseen kvalitatiiviseen tulkintatapaan nähden. Uusien laskennallisten tekstidatan analyysimenetelmien suhde etukäteen rakennettua koodausskeemaa noudattavaan sisällönanalyysiin (kuvio 1, ruutu 2) on erilainen. Aihemallianalyysin kaltaiset ohjaamattomat koneelliset menetelmät eivät ole koodaukseen perustuvan sisällönanalyysin suoria kilpailijoita – ellei olla sillä kannalla, että kaiken tekstidatan analyysin olisi hyvä olla aineistolähtöistä eikä valmiiseen koodijärjestelmään ja kategorioihin perustuvaa (DiMaggio, Nag & Blei 2013). Koodaamiseen perustuvan sisällönanalyysin kanssa suoremmin kilpailevat sen sijaan ohjatut (*supervised*) laskennalliset analyysimenetelmät, joissa kone ensin ”opetetaan” koodaamaan aineistoa halutulla tavalla, ja tämän jälkeen kone toteuttaa toivotun kaltaisen analyysin ihmiskoodaajiin verrattuna ylivoimaisella nopeudella ja luotettavuudella. Automatisoidun analyysin edut huomioon ottaen voi ennustaa, että koodaajatiimien työhön perustuva sisällönanalyysi tulee olemaan väistyvä tekstiaineistojen analyysitapa: ihmiskoneet (tutkimusapulaiset) tullaan yhä useammin korvaamaan oikeilla koneilla.

Keskustelua ja johtopäätökset

Olemme tässä artikkelissa pohtineet, mitä digitalisoituminen ja big data rinnakkaisilmiöineen merkitsevät sosiologian kannalta. Olemme tarkastelleet lähemmin uusia tekstidatan laskennallisia analyysimenetelmiä esimerkkinämme erityisesti aihemallianalyysi, koska tätä keskustelua on nähdäksemme hyvä käydä paitsi yleisellä tasolla myös konkreettisesti ja rajatusti. Lisäksi katsomme, että kiinnostavimmat ja potentiaalisesti syvällekäyvimmät muutokset sosiologian tutkimuskäytäntöjen kannalta tapahtuvat tekstiaineistojen analyysitapojen kehittymisen myötä.

On selvää, että aihemallien kaltaiset tekstidatan laskennalliset analyysimenetelmät tulevat kehittymään edelleen ja niiden käyttö yleistymään, mutta lähes yhtä varmaa on, että nämä menetelmät – samoin kuin muutkin uudet menetelmät ja koko keskustelu big datasta – tulevat herättämään kiistoja ja kohtaamaan vastarintaa yhteiskuntatieteilijöiden keskuudessa. Ajatus merkitysten mittaamisesta ja mallinnuksesta on yhteiskunta- ja ihmistieteiden ”skientististen” ja ”humanististen” traditioiden leikkauspisteessä (Mohr 1998). Refleksinomaisen vulgaarikritiikin

mukaan uudet menetelmät ja ajatus digitaalisista ihmistieteistä kaikkinensa edistävät ”pintapuolista analyysiä syvällisen ja läpäisevän tietämyksen sijaan”, koska ne ”uhraavat kompleksisuuden, spesifisyyden, kontekstin, syvyyden ja kritiikin skaalan, laajuuden, automaation, kuvailevien rakenteiden ja sen vaikutelman hinnalla, että tulkinta ei vaadi syvällistä konstekstuaalista tietoa” (Kitchin 2014, 143). Tietyssä mielessä syyte, että laskennalliset analyysitavat yksinkertaistavat ja redusoivat moninaisia kulttuurisia merkityksiä pitää tosin paikkansa; *kaikki* mittauspyrkimykset päätyvät pelkistämään todellisuutta jossain määrin (Mohr 1998, 364; myös Lee & Martin 2015). Aihemallianalyysin kaltaista menetelmää onkin helppo kritisoida esimerkiksi siitä, että se perustuu väärälle teorialle kielestä, tai että se irrottaa analysoimansa aineistot kontekstistaan. Syytös väärästä kielen teoriasta on kuitenkin kestämätön, koska menetelmän tarkoituksena ei ylimalkaan ole esittää teoriaa tai mallia kielestä vaan abstrahoitu ja pelkistetty kuvaus analysoitavan tekstikorpuksen perustavista merkitysrakenteista (Blei 2012a; Lee & Martin 2015; Mohr 1998). Olennaista on, yksinkertaistaako mallinnus dataa tavalla, joka on tulkinnallisesti järkevä, läpäisee sisäisen sekä ulkoisen validiteetin testit ja on käyttökelpoinen jatkoanalyyseille (DiMaggio, Nag & Blei 2013, 602). Lisäksi tehdyt tutkimukset osoittavat sen paradoksaalisen tuloksen, että ainakin tarpeeksi suurilla aineistoilla sanasäkki-oletus ja keskittyminen pelkkään semantiikkaan näyttävät toimivan riittävän hyvin, sillä syntaksin huomioiminen ei paranna merkittävästi tuloksia (ks. Mohr & Bogdanov 2013, 559). Syytös dekontekstualisoinnista on puolestaan kestämätön sikäli, että aineiston ja siitä tehtävien tulkintojen (empiirinen ja teoreettinen) kontekstualisointi on aina, menetelmästä riippumatta, *tutkijan* tehtävä. Näin ollen vaatimus kontekstualisoinnista pätee tietenkin myös uusien laskennallisten tekstidatan analyysimenetelmien kohdalla, mutta mitään erityisesti vain niihin pätevää tässä kritiikissä ei ole (vrt. emt., 559–560; myös Bail 2014, 477).

Aihemallien kaltaiset uudet laskennalliset menetelmät, jotka tähtäävät merkitysrakenteiden mallintamiseen, eivät tule syrjäyttämään tai korvaamaan perinteisiä, merkitysten tulkintaan lähemmin pureutuvia ”hermeneuttisempia” analyysitapoja (vrt. Mohr 1998, 364). Myös aihemalleja käyttävät tutkijat säilyvät tutkimusprosessin keskiössä arvioiden ja ohjaten prosessia ja välituloksia; kysymys ei ole vain automaattisesta ”napinpainamistieteestä”. Tutkimuksen toteuttaminen edellyttää edelleen kyseisen alan substanssitietämystä ja ekspertiisiä, sillä aineistolle asetettavien kysymyksien täytyy ohjata viime kädessä mallin valintaa, ja vain siten tulosten teoreettisesti informoitu ja empiirisen kontekstin huomioiva tulkitseminen on mahdollista (DiMaggio, Nag & Blei 2013, 603; Kitchin 2014, 103; Mohr & Bogdanov 2013, 560). Aihemallit kuitenkin täydentävät tärkeällä tavalla perinteistä ”kvalitatiivista” tulkitsevaa tekstianalyysiä paitsi laajentamalla potentiaalista aineistokokoa myös tarjoamalla välineitä tulkintojen systemaattisuuden parantamiseksi. Kvantifioivaan koodaukseen perustuvaan sisällönanalyysiin nähden puolestaan järjestys, jossa ”subjektiivinen” tulkinta ja ”objektiivinen” laskenta tehdään, muuttuu: perinteisessä sisällönanalyysissä tulkittiin ensin (koodijärjestelmää suunniteltaessa, mikä edellytti myös ulkopuolista tietoa ilmiöstä, ja niin edelleen) ja laskettiin lopuksi aineiston valmistuttua. Aihemallit kääntävät tämän järjestyksen ympäri: ensin lasketaan ja vasta sitten tulkitaan, mikä merkitsee tutkimusprosessin subjektiivisen elementin paikanvaihdosta. (Mohr & Bogdanov 2013, 560–561.)

Yhteiskuntatieteellinen keskustelu uusien laskennallisten tekstidatan analyysimenetelmien kuten aihemallien annista on vasta käynnistynyt. Menetelmillä on vielä omat, selvät rajoituksensa ja ongelmansa (esim. Bail 2014, 472). Näitä ovat

esimerkiksi se, että aihemallit olettavat, ettei sanojen järjestyksellä dokumenteissa ole väliä, samoin kuin se, ettei dokumenttien järjestyksellä ole koko aineistossa väliä. Useimmat aihemallit eivät tunnista aiheiden itsensä välisiä keskinäisiä yhteyksiä ja vaativat, että jokainen dokumentti jaetaan toisensa poissulkeviksi kategorioiksi. Aihemallien perusversio ei myöskään tunnista sitä, että sanasto aiheiden sisällä voi muuttua ajan kuluessa; lopulta aineistosta eroteltujen aihemallien määrän varmistaminen vaatii validiointia.

Kulttuurisosiologian kannalta olennainen kysymys on, voiko aihemalleja käyttää sellaisten kulttuuristen elementtien kuten ”kehysten” (Goffman) tai ”symbolisten rajojen” (Lamont) aineistolähtöiseen tunnistamiseen ja luokitteluun (Bail 2014, 465–482). Vastaukset tähän vaihtelevat. Esimerkiksi DiMaggio kumppaineineen ovat hyvin optimistisia ja näkevät aihemallien menetelmän ja kulttuurisosiologisten teorioiden välillä poikkeuksellisen hyvän yhteensopivuuden, mikä ilmenee siinä, miten aihemalleja voi käyttää kulttuurisosiologian avainkäsitteiden operationalisoimiseen empiirisesti (DiMaggio, Nag & Blei 2013, 585–593, 602). Toisaalta Bail (2014) on skeptisempi ja kyseenalaistaa esimerkiksi kehyksen käsitteen suoran tunnistettavuuden sanojen esiintymisen ja sanaryhmien mukaisesti. Oma kysymyksensä on, eroavatko erilaiset tekstityypit merkittävästi siinä, kuinka käyttökelpoinen aihemallianalyysi on. Esimerkiksi narratiivien on sanottu olevan tekstityyppi, jonka analysoimiseksi aihemallit eivät ole parhaimmillaan (Mohr & Bogdanov 2013, 559).

Aihemallien kaltaiset tekstianalyysimenetelmät ovat joka tapauksessa liian kiinnostavia ja lupaavia uusia menetelmiä, etteikö sosiologien kannattaisi ottaa niitä menetelmävalikoimaansa. Käyttöönotto ei konkreettisesti tapahdu aivan vaivattomasti, sillä vaikka analyysiohjelmat vähitellen epäilemättä kehittyvätkin helppokäyttöisimmiksi (graafiset liittymät jne.), vaatii niiden hallinta ainakin vielä nykyisellään jonkinlaista ohjelmointi- tai tietojenkäsittelytieteellistä taustaa. Onkin sanottu, että useimpien sosiologien kannalta uusien menetelmien käytön mahdollistavien taitojen opettelulla on verraten korkeat ”sisäänpääsykulut” (Bail 2014, 478). Käytännössä sosiologien tulee olla valmiita rohkeaan ja uudenlaiseen tieteenalojen rajat ylittävään yhteistyöhön tietojenkäsittelytieteiden ja muiden laskennallisen data-analyysin edustajien kanssa.

Tässä artikkelissa tarkastellut uudet tekstidatan laskennalliset analyysimenetelmät kuuluvat koneoppimisen (*machine learning*) alaan, joka on tekoälyn (*artificial intelligence*) alalaji tietojenkäsittelytieteessä. Yhteistyön ja uusien menetelmien käyttöönoton tarve voidaankin tiivistää viittaamalla tilastolliseen analyysiin viime vuosikymmeninä kehittyneeseen kahteen eri asioita painottavaan paradigmaan tai koulukuntaan: yhtäältä tilastotieteen perinteeseen, todennäköisyysteoriaan ja käytetyn aineiston estimointiin nojaavaan data-analyysin kulttuuriin, ja toisaalta tietojenkäsittelytieteeseen, algoritmeihin ja ennustamiseen nojaavaan koneoppimisen kulttuuriin (Breiman 2001; Friedman 1998). Tilastotieteen sisällä käyty keskustelu frekventistisestä ja bayesilaisesta päättelystä on tuonut kulttuureja jonkin verran lähemmäs toisiaan, mutta yhteiskuntatieteet ja sosiologia ovat tähän mennessä hyödyntäneet ja tehneet yhteistyötä oikeastaan vain ensiksi mainitun tilastollisen analyysin tradition kanssa. Olisi sangen erikoinen ja sosiologian itsensä kannalta kestämätön tilanne, jos menetelmällisiä välineitä ei jatkossa omaksuttaisi yhä enemmän myös koneoppimisen kulttuurista.

Kiitokset

Artikkeliin johtanut yhteistyö sai alkunsa Arho Toikan esitelmästä ”Computational Analysis of

Text Data" Semi Purhosen johtamassa KuHis E (Kulttuuri, hierarkia ja sosiaalinen eriytyminen) -seminaarissa Helsingin yliopistossa 9.4.2014. Kiitämme aiheesta tuolloin ja myöhemmin kanssamme keskustelleita. Artikkeli on osa Suomen Akatemian (291619) ja Koneen Säätiön rahoittamaa tutkimushanketta "Cultural Distinctions, Generations and Change: A Comparative Study of Five European Countries, 1960–2010".

Kirjallisuus

BAIL, CHRISTOPHER A. 2014. "The Cultural Environment: Measuring Culture with Big Data." *Theory and Society* 43:3/4, 465482.

BIERNACKI, RICHARD. 2012. *Reinventing Evidence in Social Inquiry: Decoding Facts and Variables*. New York: Palgrave Macmillan.

BIERNACKI, RICHARD. 2014. "Humanist Interpretation Versus Coding Text Samples." *Qualitative Sociology* 37:2, 173–188.

BLEI, DAVID M. 2012a. "Probabilistic Topic Models." *Communications of the ACM* 55:4, 77–84.

BLEI, DAVID M. 2012b. "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2:1. <http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/> (luettu 17.5.2015)

BLEI, DAVID M. 2014. "Build, Compute, Critique, Repeat: Data Analysis with Latent Variable Models." *Annual Review of Statistics and Its Application* 1, 203–232.

BLEI, DAVID M. & JOHN D. LAFFERTY. 2006. "Dynamic Topic Models." *Proceedings of the 23rd International Conference on Machine Learning*. Pittspurgh: ICML, 113–120.

BLEI, DAVID M., ANDREW Y. NG & MICHAEL I. JORDAN. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3, 993–1022.

BLOK, ANDERS & MORTEN AXEL PEDERSEN. 2014. "Complementary Social Science? Quali-Quantitative Experiments in a Big Data World." *Big Data & Society* 1:2, 1–6.

BOYD, DANAH & KATE CRAWFORD. 2012. "Critical Questions for Big Data: Provocations for a Cultural, Technological and Scholarly Phenomenon." *Information, Communication & Society* 15:5, 662–679.

BREIMAN, LEO. 2001. Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author). *Statistical Science* 16:3, 199–231.

BURROWS, ROGER & MIKE SAVAGE. 2014. "After the Crisis? Big Data and the Methodological Challenges of Empirical Sociology." *Big Data & Society* 1:1, 1–6.

CHANG, JONATHAN, JORDAN BOYD-GRABER, SEAN GERRISH, CHONG WANG & DAVID M. BLEI. 2009. "Reading Tea Leaves: How Humans Interpret Topic Models." *Advances in Neural Information Processing Systems* 22, 288–296.

DANILEVSKY, MARINA, CHI WANG, NIHIT DESAI, JINGYI GUO & JIAWEI HAN. 2013. "KERT: Automatic Extraction and Ranking of Topical Keyphrases from Content-Representative Document Titles." *http://arxiv.org/pdf/1306.0271v1.pdf* (luettu 14.1.2016)

DIMAGGIO, PAUL, MANHISH NAG & DAVID BLEI. 2013. "Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding." *Poetics* 41:6, 570–606.

DUTCHER, JENNA. 2014. "What is Big Data?" Berkeley School of Information, datascience@berkeley Blog. <http://datascience.berkeley.edu/what-is-big-data> (luettu 17.5.2015)

DUTTON, WILLIAM H. & MARK GRAHAM. 2014. "Introduction." Teoksessa *Society and the Internet: How Networks of Information and Communication are Changing Our Lives*, toim. Mark Graham & William H. Dutton. Oxford: Oxford University Press, 1–20.

EROLA, JANI & PEKKA RÄSÄNEN (TOIM.) 2014. *Johdatus sosiologian perusteisiin*. Helsinki: Gaudeamus.

FLIGSTEIN, NEIL, JONAH BRUNDAGE & MICHAEL SCHULTZ. 2014. "Why the Federal Reserve Failed to See the Financial Crisis of 2008: The Role of 'Macroeconomics' as Sense-Making and Cultural Frame." Paper presented at the Annual Meetings of the American Sociological Association, San Francisco, CA, August 16–19, 2014.

FRIEDMAN, JEROME H. 1998. "Data Mining and Statistics: What's the Connection?" *Computing Science and Statistics* 29:1, 3–9.

GEERTZ, CLIFFORD. 1973. *The Interpretation of Cultures*. New York: Basic Books.

GOLD, MATTHEW K. (TOIM.) 2012. *Debates in the Digital Humanities*. Minneapolis: Minnesota University Press.

GRIMMER, JUSTIN. 2010. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." *Political Analysis* 18:1, 1–35.

GRIMMER, JUSTIN & BRANDON M. STEWART. 2013. "Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts." *Political Analysis* 21:3, 267–297.

HEIKKINEN, VESA. 2006. "Uudenvuodenpuheiden piirteitä 1935-2006 ja näkymiä vallan medioitumiseen." Teoksessa *Kieli ja teknologia*, Toim. Tuija Nikko & Pekka Pälli. Helsinki: Helsinki School Of Economics, 173–194.

HEIKKINEN, VESA & MIKKO LOUNELA. 2008. "Small Corpus, Great Institution – And an Attempt to Understand Them." *Proceedings of the 19th European Systemic Functional Linguistics Conference and Workshop 23rd - 25th July 2007*, toim. Erich Steiner & Stella Neumann. Saarbrücken, 1–32.

HERRERA, YOSHIKO M. & BEAR F. BRAUMOELLER. 2004. "Introduction to the Symposium: Discourse and Content Analysis." *Qualitative Methods* 2:1, 15–19.

IGNATOW, GABE. 2015. "Theoretical Foundations for Digital Text Analysis." *Journal for the Theory of Social Behaviour*. Article first published online: 27 FEB 2015.

JANASIK, NINA, TIMO HONKELA & HENRIK BRUUN. 2009. "Text Mining in Qualitative Research Application of an Unsupervised Learning Method." *Organizational Research Methods* 12:3, 436–460.

JÄRVELIN, KALERVO & ARI PIRKOLA. 2005. "Morphological Processing in Mono- and Cross-Lingual Information Retrieval." Teoksessa *Inquiries into Words, Constraints and Contexts*, toim. Antti Arppe, Lauri Carlson, Krister Lindén, Jussi Piitulainen, Mickael Suominen, Martti Vainio, Hanna Westerlund & Anssi Yli-Jyrä. Stanford: CSLI Publications, 214–226.

KITCHIN, ROB. 2014. *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*. London: Sage.

LASSWELL, HAROLD D., DANIEL LERNER & ITHIEL DE SOLA POOL. 1952. *The Comparative Study of Symbols: An Introduction*. Palo Alto: Stanford University Press.

LAZER, DAVID, ALEX PENTLAND, LADA ADAMIC, SINAN ARAL, ALBERT-LASZLÓ BARABÁSI,DEVON BREWER, NICHOLAS CHRISTAKIS, NOSHIR CONTRACTOR, JAMES FOWLER, MYRON GUTMANN, TONY JEBARA, GARY KING, MICHAEL MACY, DEB ROY & MARSHALL VAN ALSTYNE. 2009. "Computational Social Science." *Science* 323:5915, 721–723.

LEE, MONICA & JOHN LEVI MARTIN. 2015. "Coding, Counting and Cultural Cartography." *American Journal of Cultural Sociology* 3:1, 1–33.

LEETARU, KALEV H. 2012. *Data Mining Methods for the Content Analyst: An Introduction to the Computational Analysis of Content*. New York: Routledge.

LEWIS, SETH C., RODRIGO ZAMITH & AFRED HERMIDA. 2013. "Content Analysis in an Era of Big Data: A Hybrid Approach to Computational and Manual Methods." *Journal of Broadcasting & Electronic Media* 57:1, 34–52.

LIGHT, RYAN. 2014. "From Words to Networks and Back: Digital Text, Computational Social Science, and the Case of Presidential Inaugural Addresses." *Social Currents* 1:2, 111–129.

LYON, DAVID. 2014. "Surveillance, Snowden, and Big Data: Capacities, Consequences, Critique." *Big Data & Society* 1:2, 1–13.

MAYER-SCHÖNBERGER, VIKTOR & KENNETH CUKIER. 2013. *Big Data: A Revolution That Will Transform How We Live, Work and Think*. London: John Murray.

MOHR, JOHN W. 1998. "Measuring Meaning Structures." *Annual Review of Sociology* 24, 345–370.

MOHR, JOHN W. & PETKO BOGDANOV. 2013. "Topic Models: What They Are and Why They Matter." *Poetics* 41:6, 545–569.

MOHR, JOHN W. & AMIN GHAZIANI. 2014. "Problems and Prospects of Measurement in the Study of Culture." *Theory and Society* 43:3/4, 225–246.

MOHR, JOHN W., ROBIN WAGNER-PACIFICI, RONALD L. BREIGER & PETKO BOGDANOV. 2013. "Graphing the Grammar of Motives in National Security Strategies: Cultural Interpretation, Automated Text Analysis and the Drama of Global Politics." *Poetics* 41:6, 670–700.

MÜTZEL, SOPHIE. 2015. "Structures of the Tasted: Restaurant Reviews in Berlin between 1995 and 2012." Teoksessa *Moments of Valuation: Exploring Sites of Dissonance*, toim. Ariane Berthoin Antal, Michael Hutter & David Stark. Oxford: Oxford University Press, 147–167.

NEUENDORF, KIMBERLEY A. 2002. *The Content Analysis Guidebook*. Thousand Oaks: Sage.

PENTLAND, ALEX. 2014. *Social Physics: How Good Ideas Spread – The Lessons from a New Science*. New York: The Penguin Press.

PHILLIPS, NELSON & CYNTHIA HARDY. 2002. *Discourse Analysis: Investigating Processes of Social Construction*. Thousand Oaks: Sage.

RUPPERT, EVELYN, JOHN LAW & MIKE SAVAGE. 2013. "Reassembling Social Science Methods: The Challenge of Digital Devices." *Theory, Culture & Society* 30:4, 22–46.

SAVAGE, MIKE & ROGER BURROWS. 2007. "The Coming Crisis of Empirical Sociology." *Sociology* 41:5, 885–899.

SAVAGE, MIKE & ROGER BURROWS. 2009. "Some Further Reflections on the Coming Crisis of Empirical Sociology." *Sociology* 43:4, 765–775.

SCHROEDER, RALPH. 2014. "Big Data: Towards a More Scientific Social Science and Humanities?" Teoksessa *Society and the Internet: How Networks of Information and Communication are Changing Our Lives*, toim. Mark Graham & William H. Dutton. Oxford: Oxford University Press, 164–176.

TANG, JIAN, ZHAOSHI MENG, XUANLONG NGUYEN, QIAOZHU MEI & MING ZHANG. 2014. "Understanding the Limiting Factors of Topic Modeling via Posterior Contraction Analysis." *Proceedings of the 31st International Conference on Machine Learning*. Beijing: ICML, 190–198.

TINATI, RAMINE, SUSAN HALFORD, LESLIE CARR & CATHERINE POPE. 2014. "Big Data: Methodological Challenges and Appproaches for Sociological Analysis." *Sociology* 48:4, 663–681.

TÖTTÖ, PERTTI. 2012. *Paljonko on paljon? Luvuilla argumentoinnista empiirisessä tutkimuksessa*. Tampere: Vastapaino.

ZAMITH, RODRIGO & SETH C. LEWIS. 2015. "Content Analysis and the Algorythmic Coder: What Computational Social Science Means for Traditional Modes of Media Analysis." *The ANNALS of the American Academy of Political and Social Science* 659:1, 307–318.