Maiju Kannisto
FM, Kulttuurihistoria, Turun yliopisto

Uusi, ehompi, paras? Digitaaliset tekstilouhinnan työkalut televisiotutkimuksessa

Viime vuosina humanistiseen tutkimukseen on rantautunut uusia digitaalisia työkaluja tekstien analysointia varten. Uudet metodit eivät suoraan tarjoa jotain parempaa, niiden mahdollisuuksia ja rajoituksia on syytä pohtia kriittisesti. Tässä katsauksessa lähestyn aihetta esitellen kahta tekstilouhinnan työkalua. Tarkastelen MALLETin ja *Voyant toolsin* käyttöä omaan aineistooni, MTV:n mediaoppaisiin. MALLET etsii yhdistäviä rakenteita tekstistä eli tekee *topic modelingia*, kun taas *Voyant toolsilla* voi itse tutkia tekstiä sanojen tasolla.

Tutkimusmetodien valinta on koko tutkimusprosessia ohjaava analyyttinen työkalu. Yhteiskuntatieteellisen tutkimuksen traditioon kuuluvassa viestinnän tutkimuksessa metodi on väline, jonka nimeämisellä määritetään paikka tieteenalalla (Keinonen 2008, 132–133). Humanistisessa tutkimuksessa tutkijan oma tulkinta on keskiössä, eikä luonnontieteellisen kokeen keskeinen kriteeri toistettavuudesta toteudu tiettyä metodia käyttämällä. Tutkijan oma tulkinta tuottaa uutta tietoa ja näkökulmaa aineistoon ainutkertaisen kokemuksen kautta. Kuitenkin tutkimuksen kysymyksenasettelu, tulkintaprosessi ja johtopäätökset on kirjoitettava esiin läpinäkyvästi ja argumentoiden. (Nivala & Mähkä 2012, 9–10.) Viime vuosina television historiantutkimuksessa yleistynyt kulttuurihistoriallisuus painottaa aineistolähtöisyyttä ja laajaa empiiristä aineistoa (Keinonen 2008, 133; Salokangas 2005).

Laajan empiirisen aineiston lisäksi pitkän keston ja jatkuvuuden tarkastelu on noussut viime vuosina humanististen tutkijoiden kiinnostuksen kohteeksi. *Big data* eli digitaalisuuden mahdollistamat suuret aineistot ja toisaalta muun muassa ympäristömuutoksen tapaiset pitkän keston ongelmat ovat nostaneet esiin kysymyksiä ja kiinnostusta muutoksesta sadan ja jopa tuhannen vuoden perspektiivissä (Guldi & Armitage 2014). Myös pyrkimyksiä tuoda luonnontieteellistä mitattavuutta humanistiseen tutkimukseen on esiintynyt aina, mutta viime vuosina vaatimukset ovat taas olleet äänekkäämpiä.

Massiivisten tietokantojen käyttö on uusi tapa hyödyntää määrällisiä metodeja tutkimuksessa. Myös televisiotutkimuksen osalta on syytä pohtia uusia metodeja tutkimukseen. Määrällistä tutkimusta on sinällään käytetty laadullisen tutkimuksen ohessa televisiotoimintaa hahmotettaessa niin tv-alan toimijoiden kuin akateemisten tutkijoidenkin tutkimuksissa. Esimerkiksi liikenne- ja viestintäministeriön toteuttamissa vuosittaisissa *Suomalainen tv-tarjonta* -tutkimuksissa tilastoidaan ohjelmatarjontaa ja Finnpanelin tv-mittaritutkimuksissa katsojalukuja. Digitaalinen tekstilouhinta tuo määrällisyyden kuitenkin laadullisen tutkimuksen ytimeen – tekstin tulkintaan.

MALLET ja *topic modeling*

Topic modeling tarjoaa sarjan algoritmeja etsimään piilotettuja temaattisia rakenteita isosta tekstiaineistosta. *Topic modelingin* tuloksia voidaan käyttää tiivistämään, visualisoimaan, tarkastelemaan tai teoretisoimaan aineistoa (Blei 2012). MALLET eli *The MAchine Learning for LanguagE Toolkit* on suosituimpia *topic modelingin* digitaalisia työkaluja humanistien keskuudessa. Se perustuu *Latent Dirichlet Allocation* (LDA) -malliin ja kehitettiin Massachusetts Amherst -yliopistossa jo vuonna 2002. Kuitenkin vasta nyt myöhemmin menetelmä on saavuttanut suurempaa mielenkiintoa. *Topic modeling* ja MALLET erityisesti ovat kiinnostaneet historioitsijoita, jotka voivat käyttää sitä laajojen arkistoaineistojen analyysiin, kunhan aineisto on digitoitu ja käynyt läpi tekstintunnistuksen (OCR). MALLET käsittelee suuria tekstimääriä muodostaen sanaklustereita eli aihioita sanoista, jotka esiintyvät usein yhdessä. Humanistisessa tutkimuksessa Robert K. Nelsonin sanomalehtiaineistoa käyttänyt *Mining the Dispatch* (2010) ja Cameron Blevinsin päiväkirja-aineistoon pohjautuva *Topic Modeling Martha Ballard's Diary* (2010) olivat varhaisimpia MALLETia soveltaneita tutkimuksia (Graham & Milligan 2013).

Nelsonin tutkimuksessa *topic modeling* mahdollistaa kaukoluennan[1] laajasta sanomalehtiaineistosta: vuosien 1860–1865 välillä julkaistusta 112 000 kirjoituksesta eli 24 miljoonasta sanasta. Kun lähiluennassa olisi pakko turvautua pieneen otokseen, tekstilouhinnassa voi tutkia ja jäljittää koko aineistossa toistuvia malleja ja rakenteita (Nelson 2010). Myös Blevins tutkii laajaa 27 vuoden päiväkirjamateriaalia, josta *topic modelingin* avulla on mahdollista tutkia temaattisia trendejä. Päiväkirja-aineiston lyhyet, sisältölähtöiset kirjoitukset tuottavat yhtenäisiä ja täsmällisiä aihioita, joita Blevins nimeää. Blevinsin mukaan MALLET ryhmittelee sanoja jopa paremmin kuin ihmislukija ja yllättää sanojen yhteyksillä (Blevins 2010). Pääsin testaamaan *topic modelingia* MALLETilla omaan aineistooni Mila Oivan pitämällä luennolla "Topic modeling – määrällislaadullista analyysia?" Turun yliopistossa 30.11.2015.

Topic modeling tehtiin MTV:n keväiden 1989–1996 *Media Manager* -oppaiden tekstisisältöön. Analysoitavaa aineistoa kootessani jouduin tekemään jo monta tutkimuksellista valintaa. Aineiston yhteneväisyyden vuoksi valitsin keväiden mediaoppaat, koska niitä löytyi joka vuodelta, kun taas jotkin tutkittavan ajanjakson kesä- ja syyskauden mediaoppaat puuttuivat aineistostani. Ohjelmakaudet eroavat toisistaan kevään ollessa strategisesti merkittävin. Poistin mediaoppaiden tekstianalyysista ohjelmatietojen osuuden, koska olin tässä yhteydessä kiinnostunut nimenomaan mainosmyynnin muutoksesta. Tutkimusajanjaksoksi valikoituivat vuodet 1989–1996, koska sinä aikana mediaoppaat sisälsivät suurin piirtein vastaavia sisältöjä vuosittain ja ne löytyivät yhtenäisesti aineistostani – toisaalta ajoitus myös osui yhteen kilpailun aikakaudeksi kutsumani ajan kanssa (Kannisto 2015). MTV:n suuntautuessa kohti omaa kanavaa vuoden 1989 kanavajakopäätöksen jälkeen ohjelmasijoittelua voitiin tehdä täysimääräisesti omalla kanavalla. Omana kanavana kilpailtiin Yleisradion kanavia, 1997 aloittanutta Nelosta ja muita medioita vastaan. Vuonna 1997 alkaneen MTV2000-kehitysprojektin myötä MTV:ssä alettiin suuntautua digitalisointiin, kanavaperheen laajentamiseen ja muihin sähköisiin medioihin.

Topic modelingin tuloksena nousi yksi aihio erityisen tärkeäksi (54 %). Vaikka sanalistaukseen mahtui mukaan useita täytesanoja, kuten "myös", "sekä" ja "että", niin aihio paljasti uutisten ja *Huomenta Suomen* erityisen suuren merkityksen mai-

[1] Termi *kaukoluenta* tulee Franco Morettilta, joka peräänkuulutti kirjallisuustieteisiin lähiluennan vastapainoksi kaukoluentaa (*distant reading*). Tällä hän viittasi historioitsija Fernand Braudelin pitkän keston näkökulmaan. (Moretti 2005.)

nosmyynnissä. *Huomenta Suomi* alkoi juuri tarkastelujaksoni ensimmäisenä vuonna 1989 ja tarjosi kokonaan uudenlaisen mediatuotteen ja mainosajan aamulla ja aamupäivällä. Aamulähetykset saavuttivat nopeasti katsojia, mistä saatiin pontta mainosmyyntiin. ”*Huomenta Suomi!* – ja tuotteesi on päivän ostoslistalla”, todettiin syksyn 1993 *Media Manager* -oppaan mediaesittelyssä. Tarkastelujakson loppupäässä 1995–1996, *Huomenta Suomen* vakiinnuttua se ei ollut aineistossa enää niin voimakkaasti esillä. Ilman *topic modelingin* tulosta en olisi osannut nostaa *Huomenta Suomen* roolia niin merkittäväksi.

Topic modelingin tulos haastoi miettimään uusia näkökulmia aineistoon menetelmän ohjatessa pohtimaan yhdistävää tekijää sille, miksi algoritmi oli yhdistänyt tietyt sanat aihioiksi. Yhdistävän tekijän pohtiminen ja aihioiden nimeäminen kuuluvatkin oleellisena osana tutkimusprosessiin, aihioiden oudoilta kuulostavat sanalistat eivät itsessään vielä ole valmiita tutkimustuloksia. *Topic modeling* ei korvaakaan lähilukua vaan tarjoaa tekstiin uusia näkökulmia ja yhteyksiä asioiden välillä. Nelsonin mukaan *topic modelingin* arvo tulee esiin silloin, kun metodin avulla päästään kiinni yhteyksiin, joita emme osaa helposti selittää ja jotka yllättävät ja johdattavat meidät siten kiinnostavien ja hyödyllisten tutkimuskysymysten äärelle (Nelson 2010). *Topic modeling* voi johdattaa tutkijan suuntaan, jonne muuten ei olisi osannut tai huomannut lähteä, ja toisaalta se saattaa nostaa esiin teemoja, joista ei olisi tutkijana muuten ollut niin kiinnostunut. Esimerkiksi ihmeekseni huomasin, kuinka lähiluennassani vähälle huomiolle jäänyt *infomercial*-mainostustapa nousi yllättäen esiin *topic modelingin* tuloksissa.

Ohjelman teknisen puolen huono ymmärrys voi hankaloittaa analyysin tekoa. Tutkija joutuu testailemaan aineistonsa kanssa, kuinka monta aihiota tuottaa hyödyllisimmän tuloksen. Omassa analyysissani kymmenen oli sopiva määrä, mutta niissäkin yksi aihio nousi merkittävimmäksi (suhdeosuus 54 %) muiden jäädessä huomattavasti pienemmiksi (4,4–6,5 %). Kuten John Graham ja Ian Milligan huomauttavat MALLETia arvioidessaan, ohjelma ei taianomaisesti tuota tutkimustuloksia, vaan tulokset voivat olla harhaanjohtavia ilman eri vaihtoehtojen testailua. He suosittelevatkin tarkkoja muistiinpanoja eri muuttujilla saaduista tuloksista osana tutkimusprosessia (Graham & Milligan 2013).

Topic modelingin tarjoamat aihiot sellaisenaan voivat tarjota valheellisen tunteen sanojen merkitysyhteydestä. MALLET ei tunnista sanojen ja ilmiöiden yhteyksiä ja voi tuottaa vääristyneitä tuloksia analysoidessaan sanat irrallaan ja yksittäisinä ilman kontekstiaan (ks. esim. Schmidt 2012). Tutkijan onkin tunnettava aineistonsa ja ennen tietokoneavusteista luentaa mietittävä mahdollisia pois suljettavia sanoja, joita voi syöttää MALLETin *stop list* -toimintoon. Suomen eri sijamuodot ovat myös haaste, koska ohjelma ei tunnista taivutettuja sanoja samaksi. Lisäksi on huomioitava virhemarginaali OCR-tekstintunnistuksessa, joka saattaa jättää tunnistamatta sanoja. Koko tekstin korjaaminen käsin olisi työläs lisävaihe, joten tutkijan on pohdittava, riittääkö tekstimassan suuri koko kompensoimaan virheet ja pitämään virhemarginaalin kohtuullisena.

Tutkimusjulkaisuissa ei yleensä käydä läpi metodologisia vaikeuksia ja epävarmuuksia, mutta ainakin varhaisissa MALLETia käyttäneissä Nelsonin ja Blevinsin tutkimuksissa tuodaan näkyviin myös *topic modelingin* haasteita ja tutkimuksen prosessia laajemmin. *Topic modelingin* tuloksia esittelevien visualisointien taakse kätkeytyy kuitenkin myös paljon näkymättömiä valintoja ja ohjelman testailua, jota tutkimusprosessi vaatii. Työkalujen vakiintuessa tutkimusprosessin esittely tulee varmasti vähenemään.

Uutuuden analyysia *Voyant toolsin* avulla

Tutkimuksellinen kiinnostukseni kohdistui siihen, miten ja minkälaisia merkityksiä mainostilalle ja kuluttajille annettiin MTV:n mainostajille suunnatussa materiaalissa ja minkälaista muutosta retoriikassa on tuolla aikavälillä. Retoriikan muutoksia aineistossani jäljitin helposti käytettävällä *Voyant tools* -työkalulla. Se on *open source* web-pohjainen sovellus, joka mahdollistaa perustoiminnot tekstilouhinnassa. Perustoiminnoista *Cirrus* näyttää sanapilven useimmin esiintyneistä sanoista, *Summary* näyttää yhteenvedon tekstin sanastosta, yksittäisistä sanoista ja niiden esiintymistiheydestä ja *Corpus Reader* näyttää tekstikorpuksen kokonaisuudessaan. *Corpus Readerin* avulla voi tehdä halutuille sanoille tarkempaa analyysia sanojen esiintymistrendeistä ja -konteksteista.

James Baker hyödynsi *Voyant toolsia* sarjakuva-analyysissa käyttäen aineistonaan British Cartoon Archiven metadataa sarjakuvista vuosilta 1960–1979. Hänen tutkimuksessaan eri teemoja edustavien sanojen esiintymistiheyden vertailu toi kiinnostavimman uuden näkökulman tutkimukseen. Sarjakuvan metadatan kautta hän pääsi analysoimaan myös epäsuoria merkityksiä ja näihin liittyvien teemojen käsittelyn muutoksia, koska ne oli kirjattu metadataan, vaikkeivät näkyneet suoraan sarjakuvien tekstissä (Baker 2013).

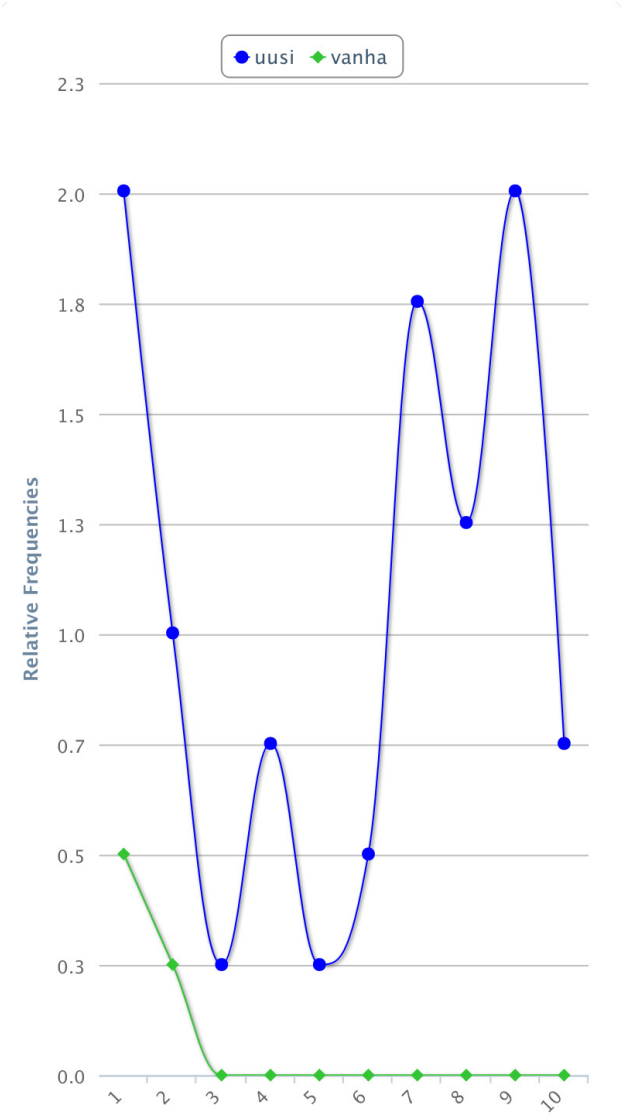
Myös omassa aineistossani jäljitin ja vertailin minua kiinnostavien sanojen esiintymistä aineistossa. Tarkastelujaksonani mainosmyynnissä tapahtui suuria muutoksia, kun kohderyhmien luokittelu tarkentui ja luokitteluilla rakennetut katsojatypologiat rakensivat katsojista mainostajille myytäviä selvärajaisia kuluttajaryhmiä suuren hahmottaman kuluttajamassan sijaan – katsojien laatu korvasi määrän. Kuluttajien ja mainostilan laadusta kiinnostuneena jäljitin adjektiiveja aineistosta. Vaikka adjektiivit eivät määrällisesti nousseetkaan käytetyimpinä sanoina esiin tekstilouhinnassa, niin retorisina keinoina ne ovat kiinnostavia tehosteita. *Voyant toolsilla* pääsin kiinni

Voyant toolsin avulla voi jäljittää käsitteiden esiintymistä laajassa aineistossa. Sanapilvi näyttää yleisimmin esiintyneet sanat, joista voi rajata pois analyysin kannalta epäoleellisia sanoja, kuten konjunktioita.

tarkempaan analyysiin käytetyistä adjektiiveista, niiden määrästä, konteksteista ja muutoksesta tutkimusjaksolla. Suomen kielen taivutusmuotojen takia eri taivutusmuotojen jäljittäminen toi oman lisätyönsä.

Suosituimpana käytettynä sanana huomio kiinnittyi ”uutuuden” korostamiseen. Uutuus oli keskeinen arvo MTV:n käyttämässä Monitor-asennetypologiassa, jossa yhdistettiin kuluttajien arvoja ja määritteitä. Keskeinen akseli oli ”uuden” etsintä, jonka vastinparina oli ”turvallisuus” 1980-luvulla ja ”tasapaino” ja ”pysyvyys” 1990-luvulla. Modernin kulutuskulttuurin voi ajatella pohjautuvan kuluttajien jatkuvaan uutuudennälkään, perustuuhan kapitalistinen talousjärjestelmä uutuuksien jatkuvaan virtaan talouden kasvun takaamiseksi.

Tarkemmin ”uusi”-sanan käytön kontekseja katsoessa tuli esiin erilaisia sille annettuja merkityksiä. Näitä merkityksiä voi hahmottaa sosiologi Colin Campbellin tekemän erottelun pohjalta. Hän erottaa toisistaan uutuuden käsitteen kolme erilaista merkitystä: 1) tuoreus vastakohtana vanhalle, 2) innovatiivisuus eli uusi paranneltu versio (viittaa erityisesti tehokkuuteen ja teknologiseen kykyyn) ja 3) modernin kulutuksen dynamiikalle keskeinen kokemuksellinen puoli (viittaa vierauden ja oudon tuomaan elämykseen, jolloin kyse on kuluttajan kokemuksesta, ei niinkään tuotteen ominaisuuksista) (Campbell 1992, 52–57).



Voyant toolsin avulla piirretty kuvaaja kertoo ”uusi” ja ”vanha” -sanojen käytön määrällisestä vaihtelusta tarkastellussa aineistossa.

Aineistossa on kuvattu ohjelmia – esimerkiksi ulkomaisia hankintoja, jotka saadaan Suomeen pian ulkomaisen ensiesityksen jälkeen – uusina ja tuoreina. Uusia mediaratkaisuja markkinoidaan paranneltuina uusina innovaatioina, joissa hyödynnetään ”viimeisintä teknologiaa”. Tällaisen markkinoinnin taustalla on edistysusko eli usko siihen, että (teknologinen) kehitys johtaa aina kohti parempaa. Uuden elämyksellinen puoli yhdistetään joihinkin ohjelmiin ja abstraktimmin mainostajille suunnattuihin uusiin mahdollisuuksiin ja vaihtoehtoihin, jotka ovat ennen kokemattomia. Eri merkitykset myös yhdistyvät. Vuonna 1995 aloittanut *Jyrki* (MTV3 1995–2001) oli Suomessa uudenlainen nuorten makasiiniohjelma, josta 1996 todettiin: ”Pitihän se arvata, että alle kolmekymppisten mielestä MTV3:n uusi JYRKI on tosi makee” (*Media Manager* 1996, 1). Ohjelma oli yhtälailla tuore, innovatiivinen ja ennen kokematon.

Uutuuteen liittyvien sanojen vastapuolena ”vanha”, ”perinteinen” ja ”tuttu” näkyivät aineistossa suhteellisen harvoin. Vuoden 1989 aineistossa esiintyi vielä sanaparit ”vanha kunnon” ja ”vanha tuttu” mutta ei enää sen jälkeen. Sen sijaan ”edullinen” ja ”tehokas” toistuivat usein, mainostajaan vedottiin näillä ja suostutteluun yhdistettiin erilaisia tutkimustietoja todisteeksi tehokkuudesta ja edullisuudesta suhteessa muihin medioihin. ”Helppo”-sana kiinnitti huomiota vuoden 1995 piikkinä. Sanojen konteksteja tarkastellessa havaitsin, että *Media Manager* -opas oli uudistunut tuolloin, ja uusia mediatuotteita ja mainonnan menestystarinoita esiteltiin kattavammin. Uudistukseen haluttiin liittää helppous käsitteenä.

”Nuoret” ja ”nuoruus” näkyivät myös paljon aineistossa kauttaaltaan. Nuoret nousivat mainostajia kiinnostavaksi kohderyhmäksi 1990-luvulle tultaessa, ja MTV halusi suunnata heille omaa ohjelmistoa. Vuonna 1991 oli piikki nuoruuden korostamisessa. Tarkemmassa analyysissa selvisi, että syksyllä 1990 tehtiin strategisia painotuksia nuorten kohderyhmään: tuolloin oli aloitettu nuorille ja nuorekkaille suunnattu populaarikulttuurin ajankohtaisohjelma *NO TV* (Kolmonen 1990–1991) ja ohjelmasijoittelussa torstain alkuilta omistettiin nuorille ja nuorille aikuisille.

Testailut *Voyant toolsilla* innostivat uusien kysymysten äärelle. Tarkemmassa lähiluennassa selvitin syitä sanojen suosion ja erityisten piikkien takana. Toisaalta *Voyant toolsilla* leikittely toimi myös vuorovaikutteisessa suhteessa osana tutkimusprosessia, kun jonkin lähiluennassa kiinnostukseni herättäneen teeman saattoi ottaa määrällisen tarkastelun kohteeksi.

Voyant toolsin käytön haasteena olivat jotkut tekniset ongelmat ohjelmassa, etenkin tulosten tallennuksessa. Kronologisesta muutoksesta kiinnostuneelle koko tekstimassan tarkastelu antoi vain suhteellisia viitteitä sanojen määrän muutoksesta ilman tarkkaa ajoituksen mahdollisuutta. Tein tarkemman analyysin lataamalla aineistoni erikseen vuosittain ohjelmaan. Sanaston yleiskuvan saaminen näytti sanapilvenä visuaalisesti hienolta, mutta lopulta sanapilven informaatio jäi melko vähäiseksi, vaikka *stop list* -toiminnolla pystyi sulkemaan ulos täytesanoja tarkastelusta. Kaiken kaikkiaan *Voyant toolsin* suurimmat hyödyt olivat itseäni kiinnostavien sanojen analyysissa ja ohjelman tarjoamissa visualisoinneissa. Näiden hyötyjen irti saaminen vaati tutkijan omaa panosta ja aktiivisuutta, jotta osasi suunnata analyysin kiinnostavaan suuntaan.

Digitaalisten työkalujen hyödyt ja rajat

Kun pohditaan digitaalisten työkalujen hyödyntämistä televisiotutkimuksessa, oleellinen kysymys on sopivan aineiston löytäminen. Mitä laajoja tekstiaineistoja televisioon liittyen on saatavilla ja mitä aineistot voivat kertoa televisiosta, audiovisuaalisesta mediumista?

Christof Schöch pohtii digitaalista aineistoa humanistisessa tutkimuksessa ja erottaa aineistotyypit kahteen: *big data* ja *smart data*. *Big data* on järjestelemätöntä, suttuista, vaihtelevaa ja määrältään suurta, kun taas *smart data* on järjesteltyä, ihmisen käsittelemää ja määrältään vähäisempää. Humanistisen tutkimuksen *big data* ei ole samanlaista kuin luonnon- tai taloustieteiden data; määrä ei humanistisissa tieteissä ole *big dataa* parhaiten määrittävä tekijä, vaan humanistisessa tekstilouhinnassa oleellisinta on juuri työkalujen tuomat uudet metodologiset näkökulmat, jolloin niitä voi hyödyntää myös pienempiin aineistoihin. Vaikka monella suunnalla on meneillään aineistojen digitalisoimishankkeita, olemme silti vielä kaukana täydellisestä kulttuurituotteiden rekisteristä, ja todella suuret aineistot humanistisessa tutkimuksessa ovat harvassa (Schöch 2013).

Omassa tutkimuksessani MTV:n tuotannon strategioista MTV:n tuottamat erilaiset sisäiset dokumentit ja materiaalit mainostajalle ja katsojille ovat olennainen osa lähdemateriaalia. *Topic modelingin* aineiston laajuudeksi eli tekstikorpuksen kooksi Megan R. Brett suosittelee vähintään satoja tai lähemmäs tuhatta dokumenttia (Brett 2012). Oma aineistoni koostui testailussani vain kahdeksasta mediaoppaasta (lähes 40 000 sanasta), joten voi pohtia, hahmottuiko sen kokoisesta aineistosta luotettavalla tavalla kaavamaisuuksia.

Televisiotuotantoon liittyviä tekstiaineistoja, joissa hyödyntää tietokoneavusteista tekstilouhintaa, voisivat olla usean vuoden ajalta suurempina kokonaisuuksina vuosikertomukset, käsikirjoitukset, ohjelmatiedot, mainostajille suunnatut mediaoppaat, yhtiön/kanavan esitteet, sisäiset lehdet ja internet-sivujen aineisto. *Lähikuvan* arkisto-teemanumerossa 1/2012 Heidi Keinonen ja Paavo Oinonen esittelivät KAVI:n, Yleisradion ja Suomen Elinkeinoelämän Keskusarkiston (ELKA) televisioarkistoja, joista löytyy jonkin verran myös tekstimuotoista aineistoa.

Radio- ja tv-arkiston Ritva-tietokantaan on tallennettu keskeisten kotimaisten radio- ja tv-kanavien ohjelmavirtaa vuodesta 2009 alkaen. Ohjelmavirran lisäksi Ritvaan on tallennettu myös tekstiaineistona teksti-tv-sivuja ja metadataa, joita voisi hyödyntää lähdeaineistoina. Ritvan metadatan lähteinä ovat KAVI:n oman luetteloinnin lisäksi Finnpanel, Venetsia-ohjelmatietojärjestelmä, elektroninen ohjelmaopas (EPG), Radiot.fi ja kanavien toimittamat tiedot. Jokaisen sivun tekstisisältöön voi kohdistaa hakuja (Leskinen 2016).

Lisäksi täydentävänä aineistona televisiotutkimuksessa käytetään usein lehdistömateriaalia. KAVI:n leikekokoelmassa on elokuva-, radio- ja televisioaiheisia lehtileikkeitä. Kokoelmaa on ryhdytty digitoimaan elokuvista alkaen, ja siitä avautunee tutkijalle kiinnostavia tutkimusmahdollisuuksia tulevaisuudessa.

Mitä tutkimuksellisia hyötyjä digitaaliset työkalut sitten voivat tarjota tutkijalle? Analyysin merkityksellisten kohtien tunnistamisessa voi saada apua tietokoneavusteisesta luennasta. Heidi Keinonen on kuvannut temaattisen lähiluvun analyysimenetelmää, jota hän kutsuu teemoittamiseksi. Kyseessä on aineistolähtöinen analyysi, jossa laajasta ja monimuotoisesta lähdeaineistosta paikannetaan tutkimusta merkityksellistäviä kohtia. Merkityksellisistä kohdista ryhmitellään eri teemoihin liittyvät asiat ja muotoillaan teemoille merkityssisällöt itsenäisesti ja suhteessa toisiin teemoihin. (Keinonen 2011, 36, 42–44.) Tietokoneavusteinen luenta voi nostaa aineistosta esiin määrällisesti merkittäviä teemoja, jotka voi sitten nostaa tarkemman analyysin kohteeksi lähiluennassa.

Tekstilouhinnan avulla voi myös jäljittää, onko omassa tutkimusaineistossa katveita jossain teemoissa, ilmaisuissa tai ilmiöissä. Esimerkiksi MTV:n mediaoppaiden analyysissa havaitsin, että ”lama”-sanaa ei käytetty mediaoppaissa, paitsi 1995 laman jo taittuessa parempaan. Silloin korostettiin, kuinka lamavuosien aikaan tv-mainontaa käyttäneet olivat menestyneet ja kuinka lamavuosina (vuosien 1992 ja 1994 välillä) MTV-konsernin markkinaosuus kaikesta mediamainonnasta nousi

12 prosentista 20 prosenttiin. Näin rakennettiin tarinaa yhtiön lamasta selviytymisestä.

Digitaaliset työkalut ovat apukeinoja tekstin tulkintaan, mutta ne eivät saa vaarantaa humanististen tutkijoiden vahvuutta kielen tulkitsijoina ja ilmiöiden kontekstualisoijina. Tekstinlouhintatyökalut auttavat analysoimaan tekstejä, mutta aktiivinen tulkinta ja ymmärtäminen ovat edelleen tutkijan itsensä tehtäviä (Blei 2012). Tutkijan on paikannettava ja kontekstualisoitava työkalujen osoittamat ilmiöt ja trendit. Tässä työssä visualisoinnit tekstilouhinnan tuloksista ovat tärkeitä, koska niiden avulla työkalujen antamat numeraaliset tulokset eli kaavamaisuudet saadaan helpommin tulkittavaan muotoon.

Lähiluvussa tutkimusaineiston kuvien analyysi on mahdollista, mutta tekstilouhinnassa se jää väistämättä sivuun. Aineiston tuntu ja materiaalisuus, jotka myös vaikuttavat tutkijan analyysiin, eivät välity koneen kautta. Omakaan aineistoni ei ole vain sanoja, vaan retorisina keinoina toimivat myös otsikot, kuvat ja taulukot, joilla mainostajaan vedotaan. Ohjelmat eivät tunnista erilaisia muotoja, kuten erilaisia tekstityyppejä ja sävyjä, tekstissä eivätkä myöskään huomioi siinä olevia otsikoita tai muita painotuksia.

Suomalaisen tutkijan näkökulmasta tekstilouhinnan työkalujen algoritmien kehittäminen suomen kielelle olisi tärkeää. Eri sijamuotojen tunnistaminen ja muiden kielellisten erityispiirteiden huomioiminen on oleellista analyysin onnistumiselle ja laajemmalle hyödyntämiselle.

Trevor Owens erottelee kaksi tapaa käyttää *topic modelingia*: 1) tarjoamaan erilaisia näkökulmia tekstin tulkintaan ja inspiroimaan tutkimusprosessissa tai 2) argumentin rakentamisen tueksi. Ensimmäisessä tapauksessa digitaalisen työkalun rooli on toimia ajatustyön osana, mutta heti kun haluaa käyttää tuloksia argumenttinsa tukena, on oltava valmis kuvaaman analyysin kulku tarkasti ja myös perustelemaan aineiston kanssa tekemänsä valinnat. Owens painottaa, että tutkijan on ymmärrettävä, mitä voi ja mitä ei voi kertoa käyttämänsä digitaalisen työkalun tulosten perusteella (Owens 2012). Vaikka Owens puhuu nimenomaan *topic modelingista*, voi periaatetta laajentaa tekstilouhintaan yleisemminkin.

Tietokoneavusteinen luenta voi antaa inspiraatiota aineistolla leikiteltäessä tai konkreettista taustatukea argumentointiin, mutta lähiluku omassa analyysissa on edelleen välttämätöntä laajempien merkitysten tulkinnassa. Tutkijan tiedonmuodostuksen prosessissa ymmärrys muodostuu aineiston kanssa vuoropuhelua käyden. Tätä vuoropuhelua voi osaltaan tehdä digitaalisten apulaisten välityksellä.

Kiitos Mila Oivalle topic modeling -luennosta ja aineistoni analyysista MALLET-ohjelmalla.

Lähteet

Baker, James (2013) "On metadata and cartoons". *British Library, Digital scholarship blog*. <http://british-library.typepad.co.uk/digital-scholarship/2013/05/on-metadata-and-cartoons.html> (linkki tarkastettu 21.1.2016).

Blei, David M. (2012) "Topic Modeling and Digital Humanities". *Journal of Digital Humanities* vol. 2:1.

Blevins, Cameron (2010) "Topic Modeling Martha Ballard's Diary". <http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/> (linkki tarkastettu 14.1.2016).

Brett, Megan R. (2012) "Topic Modeling: A Basic Introduction". *Journal of Digital Humanities* vol. 2:1.

Graham, John & Milligan, Ian (2012) "Review of MALLET, produced by Andrew Kachites McCallum". *Journal of Digital Humanities* vol. 2: 1.

Guldi, Jo & Armitage, David (2014) *The History Manifesto*. Cambridge: Cambridge University Press.

Kannisto, Maiju (2015) ”Lauantai-ilta MTV:llä 1981–2005. Kaupallisen television strategiat ohjelmasijoittelussa”. *Lähikuva* vol. 28:4, 39–66.

KAVI:n suunnittelijan Timo Leskisen sähköpostihaastattelu 8.1.2016. Haastattelu tekijän hallussa.

Keinonen, Heidi (2011) *Kamppailu yleistelevisiosta. TES-TV:n, Mainos-TV:n ja Tesvision merkitykset suomalaisessa televisiokulttuurissa 1956–1964*. Tampere: Tampere University Press.

Keinonen, Heidi (2008) ”Metodologia prosessina. Tutkimusasetelman rakentuminen televisiohostorian tutkimuksessa”. Teoksessa Heidi Keinonen, Marko Ala-Fossi & Juha Herkman (toim.) *Radio- ja televisiotutkimuksen metodologiaa*. Tampere: Tampere University Press, 121–136.

Media Manager 1989–1995 kevät, 1996, 1. MTV Oy.

Moretti, Franco (2005) *Graphs, Maps, Trees: Abstract Models for a Literary History*. Lontoo, New York: Verso.

Nelson, Robert K. (2010) “Mining the Dispatch”. <http://dsl.richmond.edu/dispatch/pages/intro> (linkki tarkastettu 14.1.2016).

Nivala, Asko & Mähkä, Rami (2012) ”Johdanto: Lähde, menetelmä, tulkinta”. Teoksessa Asko Nivala & Rami Mähkä (toim.) *Tulkinnan polkuja. Kulttuurihistorian tutkimusmenetelmiä*. Cultural History – Kulttuurihistoria 10. Turku: k&h, 7–21.

Owens, Trevor (2012) ”Discovery and Justification are Different: Notes on Science-ing the Humanities”, <http://www.trevorowens.org/2012/11/discovery-and-justification-are-different-notes-on-sciencing-the-humanities/> (linkki tarkastettu 20.1.2016).

Salokangas, Raimo (2005) ”Mediahistorian tutkimuskohdetta etsimässä”. *Historiallinen aikakauskirja* 103:4, 483–492.

Schmidt, Benjamin M. (2012) “Words Alone: Dismantling Topic Models in the Humanities”. *Journal of Digital Humanities* vol. 2:1.

Schöch, Christof (2013) “Big? Smart? Clean? Messy? Data in the Humanities”. *Journal of Digital Humanities* vol. 2:3.

Digitaaliset työkalut

MALLET, <http://mallet.cs.umass.edu/download.php>.

Voyant Tools, <http://voyant-tools.org/>.