Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: muutama kriittinen havainto

Matti Nelimarkka*

Hyväksytty julkaistavaksi Politiikka-lehdessä

Tiivistelmä

Aihemallinnus mahdollistaa laajojen tekstiaineistojen automaattisen ryhmittelyn käyttäen ohjaamatonta koneoppimista. Kiinnostus sitä kohtaan on kasvanut ja sen soveltaminen on lisääntunut yhteiskuntatieteellisessä tutkimuksessa niin Suomessa kuin maailmalla. Kuitenkin, aihemallinnus sekä muut ohjaamattoman koneoppimisen menetelmät vaativat tutijoita tekemään momia erilaisia valintoja. Esimerkiksi tutkijat joutuvat päättämään miten aineistoa esikäsitellään, mitä koneoppimismenetelmää käytetään ja *miten* sitä käytetään sekä tulkitsemaan ohjaamattoman koneeoppimisen kautta syntyneet tulokset. Aihemallinnuksessa eräs valinta koskee aiheiden määrää, josta on käyty aktiivisesti keskustelua niin koneoppimisen kuin laskennallisen yhteiskuntatieteen yhteisöissä. Artikkelin käyttäjäkoe osoittaa, että yhteiskuntatieteissä suosittu tulkinnallisuutta korostava lähestymistapa aihemäärän valintaan on epävarma. Artikkelin empiirinen esimerkki osoittaa, että aihemäärän valinta vaikuttaa aihemallinnuksesta syntyviin tulkintoihin. Tämän pohjalta artikkeli suosittaa, että (i) parametrien valinnassa käytettäisiin tilastollisia menetelmiä ja (ii) aihemallinnuksen tulokset sidotaan yhteiskuntatieteelliseen kirjallisuuden käyttäen joko teoreettisen viitekehystä tulkinnan apuna tai käytetään aihemallinnusta menetelmällisestä trianguloiden tai grounded theory-lähtöisesti. Lisäksi suositellaan, että (iii) tutkimusprosessi voisi olla avoimempaa sekä (iv) laskennallisten menetelmien soveltajien on välttämätöntä seurata kehittyvää kriittistä algoritmitutkimusta.

ENGLISH ABSTARCT

Topic modelling is one unsupervised machine learning technique and its applications for social science have increased during the recent years, both in Finland and internationally. Topic models, like many other unsupervised machine learning methods, requite input from the researchers: what parameters are used with the unsupervised machine learning models. Through an

*
Tietotekniikan laitos, Aalto-yliopisto
Tietotekniikan tutkimuslaitos HIIT, Aalto-yliopisto ja Helsingin yliopisto
matti.nelimarkka@aalto.fi

experimental research, this article demonstrates that the commonly used interpretability focused approaches can lead to different outcomes. Through two empirical cases differences – while minor – in the outcomes of the empirical analysis. Based on the experimental and the empirical study, the article recommends that (i) the choice of parameters should be done using statistical measures instead of interpretability and (ii) the results are further elaborated using social science literature or extended analysis. Furthermore, it is recommended that (iii) there should be more transparency related to applications of supervised methods and (iv) researchers applying computational methods must follow the research on critical algorithm studies.

1 Johdanto

Laadullisten sekä määrällisten menetelmien lisäksi yhteiskuntatieteissä on viime aikoina sovellettu laskennallisia *(computational)* menetelmiä. Laskennalliset menetelmät sekä digitaalisen murroksen tuomat massaaineistot (*big data*) ovat yhdessä mahdollistaneet laskennallisen yhteiskuntatieteen (*computational social science*) käyttöönoton (Cioffi-Revilla 2010, Lazer et al. 2009). Näiden menetelmien käyttö on kasvanut merkittävästi 2000-luvulla (Savage 2013). Menetelmien on myös sanottu haastavan yhteiskuntatieteen tietokäsitystä sekä menetelmällisiä lähtökohtia (Boyd & Crawford 2012, Kitchin 2014).

Laskennallisia menetelmiä voidaan soveltaa myös tekstiaineistojen analyysiin (esimerkiksi Grimmer & Stewart 2013). Tekstiaineistojen tilastolliseen ryhmittelyyn perustuva aihemallinnus (*topic model*) on herättänyt laajasti mielenkiintoa (esimerkiksi Levy & Franklin 2014, Purhonen & Toikka 2016, Laaksonen & Nelimarkka 2018, Ylä-Anttila et al. 2018). Menetelmä on kehitetty vuonna noin kymmenen vuotta sitten, mutta laaja-alaiseen käyttöön tarkoitettu tieteellinen artikkeli on kirjoitettu vuonna 2012 (Blei 2012, Blei et al. 2000). Uutuutensa takia menetelmälle ja sen soveltamiselle ei ole vakiintuneita tapoja. Vaikka artikkelit korostavat tarvetta validoida laskennallisten menetelmien tuloksia myös tekstianalyysissä, ne eivät kerro kuinka validointi tulisi tehdä (vertaa Grimmer & Stewart 2013).

Tässä artikkelissa käsittelen aihemallinnusta menetelmänä ja siihen liittyviä haasteita. Aihemallinnus tuottaa aina tutkijan päättämän määrän ryhmiä aineistosta (aiheita), oli se viisi tai sata, ja näyttää miten tietyt sanat ja tekstit kuuluvat näihin aiheisiin. Wallach et al. (2009a) argumentoi, että aihemallinnuksen mallin parametrien[1] valinta vaikuttaa tulokseen. Onkin odotettavissa siis, että myös aihemäärillä (k) tuotetut aihemallit ja niiden tulokset ovat erilaisia. Perinteinen yhteiskuntatieteellinen lähestymistapa aihemäärän valitsemiseksi on ollut kokeilla muutamia erilaisia aihemääriä ja valita niiden perusteella tutkimuskysymyksen kannalta valaisevin aihemäärä (esimerkiksi Levy & Franklin 2014, Purhonen & Toikka 2016). Tietojenkäsittelytieteessä on taas korostettu tilastollisia mittareita ja niiden soveltamista aiheiden määrän valinnassa (Griffiths & Steyvers 2004, Wallach et al. 2009b). On kuitenkin epäselvää, miten eri tutkijat valitsevat aiheita ja mikä on aihevalinnan merkitys empiirisissä tuloksissa. Tässä artikkelissa pyritään käsittelemään näitä haasteita aihemallinnuksen sekä muiden ohjaamattomien koneoppimismenetelmien kautta.

Artikkelissa käydään ensin läpi tekstiaineiston laskennallista analyysiä yleisesti sekä aihemallinnusta prosessina. Tämän jälkeen artikkeli esittää kaksi pienempää osatutkimusta, joista kummallakin on itsenäinen kirjallisuuskatsaus sekä tulosten tulkinta keskustelun muodossa. Ensimmäisessä osatutkimuksessa käsitellään valinnan haasteita tuomalla esille toisaalta tutkijoiden erot valaisevien aihemäärien valinnassa (esimerkiksi Levy & Franklin 2014, Purhonen & Toikka 2016) sekä toisaalta tutkijoiden sekä tilastollisten lähestymistapojen eroja. Toinen osatutkimuksessa käsittelee tarkemmin aihemäärän valinnan tärkeyttä empiirisessä tutkimusprosessissa. Tätä tarkastellaan soveltamalla ensimmäisen osatutkimuksen eri aihemääriä

[1] Aihemallinnuksessa on kolme hypermatametriä: α ja β säätelevät aiheiden ja sanojen ja aiheiden sekä dokumenttien ja aiheiden todennäköisyysjakaumia ja aihemäärä k.

empiiriseen tutkimuskysymykseen Suomen puoluekentän kehittymisestä. Artikkeli päättyy johtopäätöksiin siitä, mitä kaksi tutkimuskysymystä kertovat aihemallinnuksen sekä laajemmin ohjaamattomien koneoppimismenelmien soveltamisesta yhteiskuntatieteissä. Tällöin keskustelussa otetaan kantaa laskennallisten menetelmien kriittiseen tutkimuksen puolesta (vertaa esimerkiksi Savage 2013).

2 Teksti laskennallisena datana

Tekstianalyysissä käytettävät laskennalliset menetelmät voidaan pääpiirteissään jakaa kolmeen ryhmään: sanastopohjaiseen, ohjattuun sekä ohjaamattomaan koneoppimiseen (esimerkiksi Grimmer & Stewart 2013). Tämän ryhmittelyn sijaan yhteiskuntatieteissä tutumpi lähestymistapa voi olla jako "a priori-skeemaan perustuviin menetelmiin" sekä "aineistolähtöisiin menetelmiin" (Purhonen & Toikka 2016). Molemmille on vastineet niin perinteisissä yhteiskuntatieteellisissä menetelmissä sekä uusissa laskennallisissa menetelmissä. A priori-skeemoja vastaa erilaisten olemassa olevien luokittelukehikkojen soveltaminen: esimerkiksi sisällön luokittelu koodikirjaa käyttäen on tyypillinen a priori-skeema. Aineistolähtöisistä menetelmistä grounded theory lienee tunnetuin esimerkki: aineistoon tutustuminen synnyttää siihen jonkun ryhmittelyn.

Näille menetelmille voidaan myös löytää laskennalliset vastineet. *Ohjattu koneoppiminen* vastaa a priori-skeemoja korostavia lähestymistapoja. Siinä käytetään ennalta olemassa olevaa luokittelua aineistoa (mitä usein kutsutaan opetusaineistoksi) ja etsitään tästä luokitellusta aineistosta laskennallisesti piirteitä – useimmiten sanoja – jotka mahdollisimman hyvin selittävät kuulumista luokkaan. Esimerkiksi Yhdysvalloissa republikaaniedustajat voidaan erottaa noin 60% tarkkuudella demokraattiedustajista tarkastelemalla heidän puheenvuoroja edustajanhuoneessa (Yu et al. 2008). Ohjatun koneoppimisen etuna on mahdollisuus arvioida tarkkuutta erittäin tarkasti: käytössä on sekä opetusaineiston tunnettu luokitus että koneoppimisen kautta laskettu luokitus aineistolle. Näin esimerkiksi hakevat parhaita piirteitä ja muita mallien parametreja tarkastelemalla, miten näiden eri valinnat vaikuttavat tarkkuuteen. Samoin tulosten hyödyntäminen yhteiskuntatieteellisessä tutkimuksessa on ilmeistä. Tarkkuutta voidaan mitata monella tavalla, sekä koneoppimistutkijoiden parissa (ulkoinen tarkkuus (*accuracy*), saanti (*recall*) ja sisäinen tarkkuus (*precision*)) että yhteiskuntatieteilijöiden keskuudessa (esimerkiksi Cronbachin α tai Cohenin κ). Tällöin ohjatun koneoppimisen luotettavuus korostuu ennen kaikkea sille ilmoitettuun tarkkuuteen.

Aineistolähtöisiä menetelmien laskennallinen vastine on *ohjaamattomat koneoppimismenetelmät*. Ohjaamattomissa koneoppimismenetelmissä ei ennakkoon tehdä oletuksia aineiston luokittelukriteereistä, sen sijaan aineistolle etsitään laskennallisesti ryhmiä (esimerkiksi aihemallinnus tai k-means-menetelmä) tai lainalaisuuksia (esimerkiksi assosiaatiosäännöt, katso Jurek & Scime 2014). Ohjaamattomat menetelmät ovat samankaltaisia faktorianalyysin kanssa: molemmissa aineistosta pyritään löytämään aineistolähtöisesti löytämään jotain mielekästä. Vertailu faktorianalyysiin on hyödyllistä, koska se on useille yhteiskuntatietelijöille tuttu menetelmä. Lisäksi eksploratiivinen faktorianalyysi on ollut käytössä niin kauan, että menetelmää on jo tarkasteltu

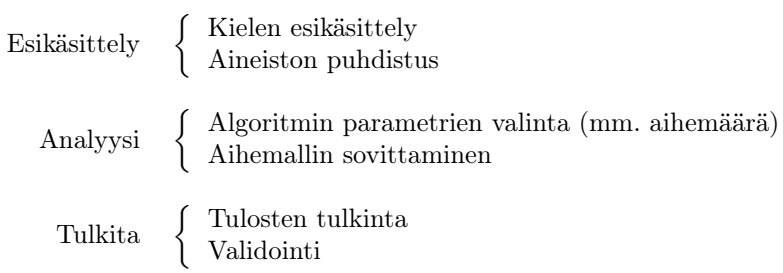
kriittisesti (esim., Fabrigar et al. 1999, Bandalos & Boehm-Kaufman 2010, Russell 2002). Psykologian aikakausilehdissä julkaistuista eksploratiivista faktorianalyysiä soveltavia artikkeleista jopa viidenneksessä analyysin tulokset on raportoitu epätarkasti ja mahdollisia virheitä esiintyy menetelmän käytössä (Fabrigar et al. 1999). Virhelähteitä ovat niin otoskoon pienuus (Fabrigar et al. 1999, Russell 2002), liian suuri faktorien määrä suhteessa aineiston kokoon (Fabrigar et al. 1999), faktorien määrän valinnan perustelut (Bandalos & Boehm-Kaufman 2010; 79–83) sekä faktorianalyysissä käytetyn menetelmän valinnat (Bandalos & Boehm-Kaufman 2010). Artikkelin fokuksessa olevien ohjaamattomien menetelmien kriittinen tarkastelu on tarpeen, koska ojaamattomaan oppimiseen perustuvat menetelmät ovat jo näyttäneet haastavuutensa.

2.1 Aihemallinnus

Aihemallinnus (*topic modelling*) ryhmittelee tekstiä (dokumentteja ja niissä olevia sanoja) ja löytää sieltä piilossa olevia rakenteita tai 'aiheita'. Matemaattisesti aihemallit perustuvat todennäköisyysjakaumiin, joiden avulla tarkastellaan dokumenttien kuulumista aiheeseen sekä sanojen kuulumista aiheisiin. Aihemallinnusta on käytetty esimerkiksi tutkiessa etujärjestöjen ja kansalaisten kommentteja lakiluonnoksiin (Levy & Franklin 2014), viestinnän painopisteitä viestintävälineiden välillä (Nelimarkka et al. arvioitavana, Laaksonen & Nelimarkka 2018) sekä presidenttien puheiden sisällön analyysiin (Purhonen & Toikka 2016) sekä ilmastomuutosta koskevien kehysten analyysissä (Ylä-Anttila et al. 2018). Aihemallinnuksessa on kolme vaihetta (kuva 1): esikäsittely, analyysi, sekä tulkinta.

2.2 Aineiston esikäsittely

Koska aihemallinnus perustuu sanojen esiintymisien laskentaan, on välttämätöntä muuttaa sanat niiden perusmuotoon. Tällöin eri sija- ja taivutusmuodoissa esiintyvät sanat muokataan samaan muotoon: esimerkiksi sanat 'kissat', 'kissoja', 'kissoille' kaikki viittaavat perusmuotoiseen asiaan 'kissa'. Kielitieteessä on kaksi lähestymistapaa perusmuotoistamiseen. Stemmauksessa sanat katkaistaan kieliopillisten sääntöjen pohjalta katkaistuun sanamuotoon. Lemmaus taas etsii sanalle perusmuotoista ilmaisua sanakirjojen sekä kielellisen analyysin sekä kautta. Näillä kahdella tavalla



Kuva 1: Aihemallinnuksen prosessikaavio

Taulukko 1: Stemmauksen ja lemmauksen eroja

	Esimerkki 1	Esimerkki 2
Alkuperäinen	Kissalla on pitkät viikset	Eilen oli vaalit
Stemmaus	kis on pitk viiks	eile oli vaali
Lemmaus	kissa olla pitkä viiksi	eilen olla vaalia

on eroja lopputuloksissa, kuten taulukossa 1 esitetään. Erityisesti stemmauksessa usein sanat voivat olla vaikeatulkintaisia. Sekä stemmaukseen että lemmaukseen löytyy monia valmiita ratkaisuja. Stemmauksessa Natural Language Toolkit (NLTK)[2] on yleisesti käytetty ratkaisu. Lemmaus vaatii taas kieltä ymmärtävän morfologista jäsentimen. Suomeksi tällainen on käytettävissä esimerkiksi Kielipankin[3] kautta.

Kielen esikäsittelyn lisäksi on välttämätöntä "puhdistaa" aineistoa jatkokäsittelyyn. Esimerkiksi teksti muutetaan pieneen kirjainkokoon ja välimerkit sekä numerot poistetaan. Lisäksi analyysistä on usein hyödyllistä poistaa yleisiä sanoja *(stopwords)*, kuten konjunktioita 'ja', 'tai' sekä 'myös' ja yleisiä verbejä 'olla'. Samalla tavoin tässä vaiheessa voidaan poistaa sanoja, jotka eivät selkeytä tulkintaa ja ovat niin yleisiä, että jopa häiritsevät tulkintaa. Toisaalta, aineistoa voi puhdistaa laajemminkin, esimerkiksi poistaa kaikki partitiivit tai erisnimet.

Tällä hetkellä ei ole olemassa selkeää ohjetta aineiston puhdistamiseen. Sen sijaan aineiston puhdistamisessa voidaan tehdä perustellusti hyvinkin erilaisia päätöksiä. Siksi onkin välttämätöntä dokumentoida tehdyt valinnat selkeästi (vertaa Denny & Spirling 2018). Aihemallinnuksen (ja ohjaamattomien koneoppimismenetelmien) osalta menetelmäkeskustelu on kuitenkin vielä lapsen kengissä. Esimerkiksi, perusmuotoistamista on pidetty kriittisenä perinteisissä menetelmäkuvauksissa, mutta tuorein kirjallisuus on haastanut tämän esikäsittelyvaiheen merkityksen: kuten Schofield & Mimno (2016) huomauttaa, stemmaus ja lemmaus voivat heikentää aihemallinnuksen laatua englannin kielellä. Toisaalta erilaisia vaihtoehtoja aineistojen puhdistamiseen on valtavasti ja niiden voidaan parantaa tai heikentää ohjaamattoman koneoppimisen tuloksia (Denny & Spirling 2018). Vaikka puhdistamiseen liittyvä menetelmäkeskustelu on vilkastunut, valitettavasti sitä käydään lähinnä englannin kielen osalta. Esimerkiksi Schofield & Mimno (2016) havainnot eivät ole yleistettävissä suomeen: englannissa persoonapäätteitä ja sijapäätteitä ei käytetä samalla tavalla kuin suomessa. Siksi suomenkielisen aineiston käsittelyssä on vielä syytä noudattaa perinteisiä ohjeita aineiston puhdistamisesta: aineisto kannattaa lemmata ennen aihemallinnusta sekä yleisiä sekä harvinaisia sanoja poistaa ennen aihemallinnusta.

2.3 Aineiston analyysi aihemallinnuksella

Kuten yllä jo kerrottiin, aihemallinnus ryhmittelee tekstiä sanojen perusteella ryhmiin eli 'aiheisiin.' Täsmällisemmin, aihemallinnus laskee jokaiselle tekstin

[2] `http://www.nltk.org/api/nltk.stem.html`
[3] `https://www.kielipankki.fi/tyokalut/`

analyysiyksikölle (eli dokumentille) aiheiden jakauman kyseisessä dokumentissa sekä jokaiselle sanalle jakauman kuulua jokaiseen aiheeseen. Jokaiselle dokumentille ja jokaiselle sanalle siis lasketaan kuuluvuus kaikkiin aiheisiin, ja kukin näistä vaihtelee välillä 0–1.

Aihemallinnusalgoritmi

Dokumenttien ja sanojen jakaumat aiheille lasketaan todennäköisyyspohjaisesti. Tällä hetkellä yleinen lähestymistapa aihemallinnukseen on LDA-menetelmä (Latent Dirichlet Allocation Blei et al. 2000, Blei 2012). Dirichlet-jakauma on siis tilastollinen jakauma, kuten yhteiskuntatietelijöille tutummat normaali- ja χ^2-jakaumat. Todennäköisyyspohjainen malli käyttää kolme hyperparametriä: alkuarvaus aiheiden jakautumisesta dokumenteille (α), alkuarvaus aiheiden jakautumisesta sanoille (β) sekä aiheiden määrä (k). Käytännössä parametrit α ja β vaikuttavat siihen, kuinka herkästi aihemallinnus tulkitsee aiheita esiintyvän eri dokumenteissa ja sanoissa.

Aihemallinnus on laskennallisesti vaativa lähestymistapa: sekä dokumenttien että sanojen jakaumia "päivitetään" aineistoa läpikäymällä (katso esimerkiksi Blei et al. 2000). Käyttäen Dirichlet-jakaumia jokaiselle sanalle ja dokumentille arvioidaan mahdollinen jakauma, että ne kuuluvat tiettyyn aiheeseen (sanakohtainen jakauma θ ja dokumenttipohjainen jakauma ϕ pohjautuvat α, β arvoihin). Näitä jakaumia parannetaan käymällä läpi jokainen aineiston dokumentti ja jokainen sen sana. Sekä sanoille että dokumenteille lasketaan Bayes-päättelyllä jatkuvasti uusia θ- sekä ϕ- jakaumia. Eli jokaisen dokumentin jokaisen sanan perustella θ- sekä ϕ- jakaumia "päivitetään" sanan ja dokumentin muodostaman yhdistelmän luoman uuden havainnon avulla. Prosessia jatketaan, kunnes jokainen dokumentti ja jokainen sana on käyty läpi. Näin kiinnostavat lopulliset arvot θ sekä ϕ perustuvat koko aineistoon ja siellä havaittuihin yhteyksiin sanoille esiintyä usein yhdessä toistensa kanssa samassa dokumentissa.

Aiheiden määrän valinta

Parametrien α ja β lisäksi aihemallinnuksessa tulee valita aiheiden määrä (k). Kyseessä on haastava vaihe tutkimusprosessissa; esimerkiksi Greene et al. (2014) nostavat aiheiden määrän valinnan keskeiseksi haasteeksi aihemallinnuksessa. Yhteiskuntatieteilijät ovat yleisesti tarkastelleet muutamia eri aihemääriä ja valinneet näistä selkeiten tulkittavan (esimerkiksi, Purhonen & Toikka 2016, Levy & Franklin 2014). Esimerkiksi Levy & Franklin (2014) perustelee, että tutkimusprosessissa keskeinen vaihe on tulkita aihemallinnuksen tuloksia. Tämän takia he valitsevat heidän mielestään selkeimpiä aiheita tuottavan aihemäärän. Toisaalta, on olemassa myös tilastollisia mittarisuureita, joiden pohjalta löydettyjen aiheiden soveltuvuutta voidaan arvioida. Silloin tutkijat voivat sovittaa aihemallin usealle eri aihemäärillä ja käyttää näitä tilastollisia suureita päättämään, mikä on paras aihemäärä. (esimerkiksi, Griffiths & Steyvers 2004, Wallach et al. 2009b). Kuten a priori-skeemojen arvioinnissa, aihemallinnuksen sopivuuden mittaamisessa on olemassa useita mittareita, kuten:

Perpleksiteetti mittaa mallin sopivuutta aineistoon laskemalla, kuinka usein tuotettu malli loisi alkuperäisen aineiston mukaisia dokumentteja. Tällöin siis pienemmät perpleksiteetti-arvot kuvaavat tilannetta, jossa malli on selkeämpi (Blei 2012, Blei et al. 2000).

Suurimman uskottavuuden harmonisen keskiarvon malli perustuu aihemallinnuksen loglikelihood-arvon tulkintaan. Loglikelihood-arvo mittaa mallin ja aineiston välistä sopivuutta. Harmoninen keskiarvo parantaa arvoa verrattuna yksittäiseen loglikelihood-arvoon (Griffiths & Steyvers 2004, Wallach et al. 2009b). Menetelmää on sovellettu useissa lähteissä, koska se yksinkertainen toteuttaa ja varsin nopea (esimerkiksi Griffiths & Steyvers 2004).

Tietojenkäsittelytieteilijät ovat esittäneet myös tarkempia mittareita aihemallinnuksen tarkkuuden arvioimiseksi. Esimerkiksi Wallach et al. (2009b) huomioivat, että Chib-estimaattori sekä vasemmalta-oikealle menetelmä voivat tuottaa tarkempia tuloksia. Molemmat menetelmät perustuvat aihemallinnusjakauman arviointiin suhteessa muihin estimoituihin aihemallinnusjakaumiin ja niiden välisen muutoksen tarkasteluun. Näiden käsittely tarkemmin ei ole vielä kuitenkaan tarpeen, koska ne puuttuvat useista valmiista aihemallinnuskirjastoista. Esimerkit kuitenkin osoittavat aihemallinnuksen olevan vielä kehittyvä ala, jonka käytännöt mukautuvat ja muuttuvat. Tämän takia soveltajien on välttämätöntä seurata menetelmäkehityksestä käytävää keskustelua aktiivisesti.

Vaikkakin laskennalliset mittarit selkeyttävät aiheiden määrän valintaa, eivät ne tietenkään ole ongelmattomia. Chang et al. (2009) osoittavat, ettei aihemallin muodostamat ryhmät ole aina täysin selkeitä. Käyttäjäkokeessaan he pyysi osallistujia jatkamaan aihemallin sanalistaa viiden sanan jälkeen ja arvioi tuloksia suhteessa aihemallin oikeisiin sanoihin. Heidän tuloksensa näyttivät, etteivät osallistujat pystyneet tulkitsemaan tilastollisesti mitattuna parhainta mallia. He päättelivät, että tilastollisen mittarin kannalta parhaan aihemallin aiheet eivät muodostaneet selkeitä kokonaisuuksia[4].

2.4 Aihemallinnuksen tulosten tulkinta

Viimeinen vaihe aihemallinnuksessa on aihemallin tulosten tulkinta sekä validointi. Aihemallinnusprosessin päätteenä yleisesti tarkastellaan sanalistoja aiheittain ja sanaryhmän perusteella aihe nimetään mielekkäästi (katso esimerkiksi taulukko L1). Sanalistojen on näytetyt tuottavan ihmisten tulkinnalle mielekkäimpiä kokonaisuuksia useista erilaisista lähestymistavoista aiheiden nimeämiseen (Aletras et al. 2017). Nimeämisprosessi on Purhonen & Toikka (2016) mukaan samankaltainen kuin faktorianalyysissä. Nimeämiselle ole yksikäsitteistä sääntöä – lopputulos on tutkijan tulkinta sanojen ja aiheiden merkityksistä. Usein ne perustuvat aiheiden yleisten tai kuvaavien sanojen käyttöön ja niiden tulkintaan.

Sanalistojen tarkastelun lisäksi on tekstianalyysissä välttämätöntä validoida tuloksia (esimerkiksi Grimmer & Stewart 2013). Vaikkakin validoinnin merkitystä

[4] Itse kyseenalaistan tämän artikkelin pohjalta yleisesti tehtyä tulkintaa välttää tilastollisia menetelmia aihemäärän valinnassa. Myös ihmisten tulkinnassa voidaan selkeästi päätyä tilanteeseen, jossa eri osallistujien mielestä selkein tulkinta on varsin erilainen.

korostetaan osana tutkimusprosessia, siihen ei tällä hetkellä tarjota kovinkaan selkeää mallia. Myöskin kirjallisuus validoinnista on varsin heikkoa. Myös tässä artikkelissa pääpaino on aihemallinnuksen suorittamisessa, eikä tuloksia pyritä tarkemmin validoimaan. Esitän kuitenkin kolme toisistaan poikkeavaa tapaa tehdä validointia käytännössä, vaikkakin alan kehityksen myötä uskon yleisesti hyväksyttyjen validointimenetelmien nousevan esille.

Yksinkertaisin tapa on tulkita aihemallinnuksen aiheita koodikirjana aineistolle. Osa aineistosta uudelleen koodattaisiin uudelleen manuaalisesti käyttäen tätä koodikirjaa. Tämän jälkeen käytettäisiin perinteisiä välineitä tutkijoiden välisen luotettavuuden arviointiin. Lähestymistavan hyvänä puolena on yksinkertaisuus sekä tuttuus kvalitatiiviselle tutkimusyhteisölle. Haasteen voi muodostaa työmäärä, koska aineistot voivat olla erittäin laajoja.

Monimutkaisempi tapa pohjautuu erilaisiin käyttäjäkokeisiin, joilla pyritään analysoimaan ihmisten ymmärrystä aiheista ja vertaamaan niitä aihemallinnuksen kautta syntyneisiin aiheisiin (Towne et al. 2016, Chang et al. 2009). Esimerkiksi Chang et al. (2009) käytti viittä yleisintä sanaa luodakseen mielikuvan siitä, mistä aiheessa on kyse. Tämän jälkeen esitettäisiin useampaa vaihtoehtoa kuudenneksi sanaksi ja voidaan suoraan arvioida, vastaako aihemallinnus ihmisen tulkintaa. Vaihtoehtoisesti voidaan näyttää kolme eri dokumenttia: kaksi samasta aiheesta ja kolmas muista aiheista (Towne et al. 2016). Jälleen on helppo arvioida, vastaako aiheet ihmisten mielekästä tulkintaa. Lähestymistapa on nopeahko suorittaa ja toimii hyvänä indikaattorina aiheiden luotettavuudesta. Valitettavasti menetelmä ei ole vakiintunut ja vaatisi tarkempaa perustutkimusta luotettavuuden osalta.

Viimeinen mahdollinen tapa on laadullisesta tutkimuksesta tuttu useamman tulkitsijan käyttö sekä heidän välinen keskustelu mahdollisista tulkinnoista. Tällöin useampi tutkija tarkastelee sanalistoja sekä muodostaa näille sanalistoille tulkintansa. Tämän jälkeen tulkinnoista keskustellaan sekä niitä täsmennetään kattamaan eri tutkijoiden näkökulmia. Lähestymistapa on jälleen yksinkertainen sekä tuttu laadulliselle tutkimusyhteisölle. Haasteena voi olla mitattavuuden puute, mikä laskennallisia menetelmiä soveltavalle yhteisölle voi olla keskeistä. Tällöin lähestymistapa voi saada osakseen kritiikkiä epämääräisyydestään.

Mitä aihemalli tuottaa?

Tulkinnan lisäksi on syytä mietitä, mitä aihemallinnuksen löytämän ‘aiheet’ ovat. Teknisestihän tuloksena on jokaiselle dokumentille jakauma eri aiheista dokumentissa sekä sanoille niiden jakauma eri aiheisiin. Tällä tavoin voidaan määrittää esimerkiksi se, mihin aiheisiin kukin dokumentti kuuluu eniten. Kuitenkin, kyseessä on varsin abstrakti määritelmä – aihemallinnus tuottaa jotain yhteen liittyviä sanaryhmittymiä ja dokumenttien ja aiheiden suhteita. Tämän takia aiheille onkin haettu monia teoreettisia vastinpareja: niiden on ajateltu olevan kehyksiä (*frames*), asioita (*issues*), teemoja (*theme*), diskursseja (*discourses*) tai draaman näytöksiä (*dramatistic scene*) (Jacobi et al. 2016, Mohr & Bogdanov 2013).

Selvää on, että keskustelu aihemallin soveltuvuudesta ja yhdistettävyydestä

Taulukko 2: Aineistona käytetyt puolueohjelmat.

Puolueen nimi	Määrä
KD	16
KESK	24
KOK	23
KOM	5
RKP	10
SDP	10
SKDL	5
SKP	10
VAS	9
VIHR	11
Muut	71

(a) Aineisto puolueittain. Kaikki puolueet joilla aineistossa vähemmän kuin viisi ohjelmaa sijoitettu muut-kategoriaan.

Vuodet	Määrä
1880–89	1
1890–99	2
1900–09	8
1910–19	7
1920–29	8
1930–39	11
1940–49	9
1950–59	8
1960–69	17
1970–79	18
1980–89	21
1990–99	37
2000–09	37
2010	5

(b) Aineisto vuosittain.

yhteiskuntatieteelliseen kirjallisuuteen tulee jatkumaan (esimerkiksi Ylä-Anttila et al. 2018). Aihemallinnus laskennallisena suorituksena on samanlainen riippumatta mitä tulkintoja ja teoreettisia käsityksiä aiheille annetaan. Sen sijaan teoreettisen merkityksen anto aiheille on uskoakseni parhaiten sidottavissa tutkimuksen taustakirjallisuuteen ja sen käsitteellistyksiin. Esimerkiksi uutismedian tutkimuksessa kehykset ovat niin vakiintunut käsite, että aihemallinnuksen tulokset pyritään tulkitsemaan kehyksinä. Selvää on, ettei aihemallinnus tuota mitään käsitettä valmiina – tähän ohjatut menetelmät ovat usein parempia, koska niissä opetetaan aineiston avulla teoreettisesti mielekäs tulkinta. Teoreettisiin käsitteisiin sitouttamattomuus on tärkeää pitää mielessä aiheiden analyysissä ja varoa liiallista argumentaatiota teoreettisten käsitteiden kautta.

3 Osatutkimuksissa käytettävä aineisto

Yksinkertaisuuden vuoksi molemmissa osatutkimuksessa käytetään samaa empiiristä aineistoa: suomalaisten puolueiden yleisohjelmia 1880-luvulta tähän päivään. Puolueohjelmat kerättiin Poliittisten ohjelmien tietovarannosta (POHTIVA-tietokanta[5]). Yhteensä aineistoon kuului 198 puolueohjelmaa, joiden jakauma ajallisesti sekä puolueittain on esitetty tarkemmin taulukossa 2. Aineisto esikäsiteltiin lemmaamalla sekä poistamalla yleiset sekä harvinaiset sanat.

[5] `http://www.fsd.uta.fi/pohtiva/`

4 Osatutkimus 1: Aihemäärän valinnan haasteet

Aihemallinnus on eräs laskennallisen analyysin keino tekstianalyysin suorittamiseen. Menetelmistä puhutaan usein pätevyyden ja uskottavuuden kautta. Pätevyys (validiteetti) liittyy tutkimuksen tekemiseen, eritoten siihen, että ilmiötä on kuvattu järkevästi. Uskottavuus (reliabiliteetti) taas liittyy siihen, että tutkimus on tehty johdonmukaisesti.

Määrällisessä tutkimuksessa pätevyyden keskeinen paino on onnistuneessa operationalisoinnissa ja otannassa: pätevässä tutkimuksessa on tarpeen mitata oikeaa asiaa sekä onnistua väitteiden yleistettävyydessä. Vastaavasti uskottavuus viittaa mittareiden kykyyn tuottaa samanlaisia tuloksia riippumatta esimerkiksi mittaajasta riippumatta (Metsämuuronen 2003). Laadullisessa tutkimuksessa käsitteet pätevyys ja uskottavuus ovat herättäneet keskustelua ja muita käsitteitä on ehdotettu. Kuitenkin, myös laadullisessa tutkimuksessakin pyrkimyksenä on saavuttaa mielekäs kuvaus tutkimuksen kohteesta. Tutkimuksen tulisi vakuuttaa lukija analyysin ja johtopäätösten mielekkyydestä, esimerkiksi tuomalla esille tutkimusaineistoa tai kuvaamalla tulkitsijan lähestymistapoja ilmiöön. Pätevyydestä ja uskottavuudesta laadullisten ja määrällisten menetelmien osalta on kirjoitettu erittäin laajasti, eikä yllä oleva lyhyt kuvaus tee oikeutta tälle varsin laajalle keskustelulle.

Kuinka pätevyyttä ja uskottavuutta tulisi käsitellä ohjaamattomissa koneoppimismenetelmissä, kuten aihemallinnuksessa? Erityisesti, jos käytetään tulkinnallisuutta korostavaa aihemäärää, kuinka tulisi huomioida tutkijoiden erilaiset lähestymistavat sekä kokemukset tutkijoina vaikuttavat siihen, mitä pidetään parhaiten tulkittavana aihemääränä?

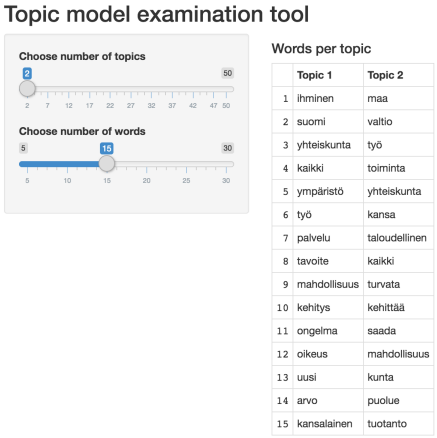
4.1 Menetelmä

Pyysimme neljää käyttäjäkokeeseen[6] osallistujaa valitsemaan aihemäärän, jolla aineistosta muodostuu selkein puoluekentää ja sen muutosta 1900-luvun alusta kuvaava kokonaisuus. Osallistujille esitetty, edellä mainittu kysymys on sellainen, että aihemallinnusta voitaisiin käyttää tämänkaltaiseen tutkimusongelmaan yhteiskuntatieteessä. Se pyrkii olemaan mahdollisimman samankaltainen yhteiskuntatieteellisessä tutkimuksessa käytetyn aihemallinnuksen sovelluksiin, vaikkakin aihemallinnuksen aiheille ei ole pyritty antamaan täsmällisempää teoriasta ja käsitteistä juontuvaa merkitystä. Kuten kuvasimme yllä, tämä vaihe usein seuraa aihemallinnusta valittujen aiheiden analyysin jälkeen.

Osallistujilla oli käytössä vuorovaikutteinen visualisaatio-ohjelma[7]. Kuten kuva 2 näyttää, osallistujat pystyivät vuorovaikutteisesti vaihtavan aihemäärää sekä tutkimaan kunkin aihemäärän 15 yleisintä sanaa. Sanalistat valittiin aiheiden

[6] Tietotekniikan tutkimuksessa asetelmia, missä henkilö käyttää ohjelmistoa tutkimusta varten kutsutaan käyttäjäkokeiksi (user study). Tätä ei pidä sekoittaa yhteiskuntatieteessä yleisemmin käytettyyn satunnaiseen koeasetelmaan (randomized controlled trial) tai luonnolliseen kokeeseen (natural experiment).

[7] Visualisaatio-ohjelma on saatavissa osoitteessa `https://github.com/HIIT/topicmodel-viz`.



	Topic 1	Topic 2
1	ihminen	maa
2	suomi	valtio
3	yhteiskunta	työ
4	kaikki	toiminta
5	ympäristö	yhteiskunta
6	työ	kansa
7	palvelu	taloudellinen
8	tavoite	kaikki
9	mahdollisuus	turvata
10	kehitys	kehittää
11	ongelma	saada
12	oikeus	mahdollisuus
13	uusi	kunta
14	arvo	puolue
15	kansalainen	tuotanto

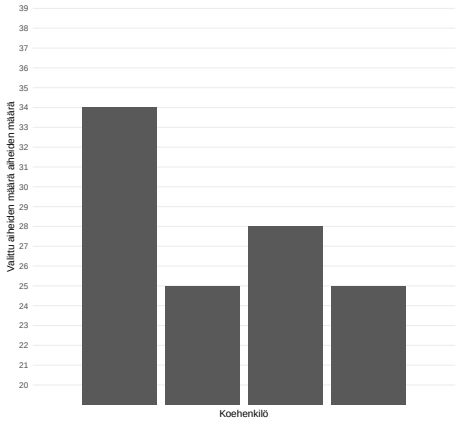
Kuva 2: Visualisointiohjelma. Vasemmalla olevasta valikosta voidaan valita aihemäärä ja ohjelma näyttää tälle aihemäärälle ja jokaiselle aiheelle 15 kuvaavinta sanaa.

esittelymuodoksi, koska niiden on havaittu tuottavan selkeimpiä ja johdonmukaisimpia tulkintoja verrattuna esimerkiksi vain tekstiesimerkkien käsittelyyn (Aletras et al. 2017). Vaikka aihemallinnus suoritettiin laajalle aihemäärälle (k=2–300), käyttäjäkokeesen valittiin vain 20 eri aihemäärää tarkasteltavaksi (k=20–39). Pienemmällä aihemäärällä haluttiin rajata osallistumiseen kuluva aika sekä vähentää käyttäjäkokeen kognitiivista kuormaa osallistujille. Aihemäärän väli valittiin kirjallisuudessa suositellun harmonisen loglikelihood-mittarin mukaan, siten että harmonisen loglikelihood-mittarin mukainen aihemäärä kuului tarkasteluvälille.

Tarkistaaksemme osallistujien ymmärryksen aihemallinnuksesta pyysimme osallistujia tutustumaan menetelmään yleistajuisen englanninkielisen artikkelin avulla (Blei 2012). Artikkelin ja menetelmän sisäistäminen tarkastettiin kolmella yleistajuisella väitelauseella. Kaikki käyttäjäkokeeseen osallistujat osoittivat hallitsevansa aihemallinnuksen perusteen. Lisäksi osallistujat olivat yhteiskuntatieteen alalta vähintään ylemmän korkeakoulututkinnon suorittaneita sekä olleet tutkijakoulutuksessa. Käyttäjäkokeeseen osallistujat siis olisivat voineet tehdä vastaavilla menetelmillä ja aineistoilla analyysiä oman tutkimuksensa osalta. Tietenkin, on ilmeistä, että kukin koehenkilö lähestyy toisaalta aineiston tulkintaa ja toisaalta avoimeksi jätettyä tehtävänannon kysymystä omien kokemuksiensa ja näkökulmansa kautta[8].

Tämän takia pyysimme osallistujia sekä valitsemaan parhaimman aihemäärän että perustelemaan, kuinka he päätyivät ehdottamaansa aihemäärään. Näitä vastauksia analysoidaan kevyen laadullisesti; koska osallistujia on vain neljä, ei ole mielekästä pyrkiä tarkkaan analyysiin. Kuitenkin perusteluiden kautta

[8] Kuitenkaan aihemallinnuksen tutkimustraditiossa ei ole toistaiseksi tapana tuoda esille tämän kaltaista subjektiivisuutta tai laajemmin reflektoida tutkijan omaa asemaa analyysiprosessissa. Noudattaen tätä toimintatapaa, emme tarkemmin käsittele osallistujien ennakkokäsityksiä tai taustaa tarkemmin tässä artikkelissa.



Kuva 3: Koehenkilöiden mielestä sopivimmat aihemäärät

voidaan kuvata myös sitä, kuinka joku aihemäärä on koettu kysymykseen sopivaksi ja mitä strategioita osallistujat käyttivät päätöksenteon tukena.

Sisällöllistä tulkintaa korostavan menetelmän lisäksi voidaan käyttää myös tilastollisia menetelmiä aihemäärän valintaan, kuten yllä esitetty. Käytämme näitä menetelmiä ja arvioimme tällä perusteella parhaimman aihemäärän sekä vertaamme sitä tutkijoiden ehdottamiin sisällöllisiin aihemääriin.

4.2 Tulokset

Osallistujien mielestä sopivin aiheiden määrä oli välillä 25 ja 34, kuten kuvasta 3 näkyy. Aihemäärien jakauma ei ole painottunut, vaan enemmänkin satunnainen. Yllättäen kaksi neljästä osallistujasta päätyivät suosittamaan samaa aihemäärää, 25 aihetta. Kaikin puolin yksinkertainen käyttäjäkoe nostaa esille jo haasteita tulkintaa korostavissa aihemäärän valinnoissa: kuinka tämä subjektiivinen valinta tulisi perustella ja mitä merkitystä aihemäärän valinnalla on tutkimuksen reliabiliteetille?

Ymmärtääksemme tarkemmin, mikä voisi selittää eroja aihemäärissä, kysyimme osallistujilta heidän kriteereistä ja strategiasta aiheen valintaan. Tunnistimme, että osallistujat käyttivät kahta eri strategiaa sopivan aihemäärän valitseminen; aihepiiristä olemassa olevan tietoa ja ennakkokäsitystä korostavaa tai aihemallinnuksen selkeyttä korostavaa. Tarkastelemme myös erikseen samaan aihemäärään päätyneiden osallistujien perusteluita.

Ensimmäinen strategia oli käyttää olemassa olevia ennakkokäsityksiä aiheiden järkevyyden ja mielekkyyden tarkasteluun. Tätä strategiaa käyttämällä aiheille annettiin poliittisesti mielenkiintoiset tulkinnat ja niitä käytettiin apuna arvioidessa aihemalleja. Esimerkiksi osallistuja kuvasi, kuinka

> Maalaisliitolla oli mielestäni esimerkiksi enemmän maatalouteen ja sitten aluepolitiikkaan ja pienyrittäjyyteen liittyvät kaksi topiikkia.

> Maalaisliitto toimi nyt tässä ikään kuin proxyna.

Tätä käyttäen hän valitsi sellaisen aihemäärän, joka toi esille niin Maalaisliiton eri poliittiset painoalueet kuin myös muita puolueita. Osallistuja päätyi 34 aiheeseen. Samoin saatettiin korostaa tiettyjen puolueiden tai politiikka-aiheiden puuttumista analyysistä ja arvioida aihemalleja, sillä perusteella, mitä ne paljastavat ilmiöstä.

> Pienemmällä aihemäärällä joukosta vaikuttaa puuttuvan joitain aiheita, jotka vaikuttavat relevanteilta, esim. Piraattipuolue-aihe.

Toinen strategia taas korosti aiheiden vertailua keskenään ja ylimääräisten tai päällekkäisten aiheiden välttämistä.

> Suuremmalla määrällä taas mukana alkaa olla epärelevantin näköisiä aiheita sekä keskenään samaa asiaa koskevia aiheita.

> Tuntui että 39 erotteli jossain määrin turhankin tarkasti tuottamalla esim. monta omaa topicia tietyille puolueille.

> Noin 25 tienoille asti tuntuu, että topiikit pysyvät selkeämmin erillisinä, sen jälkeen alkaa tulla päällekkäisyyttä ja on vaikeampi tulkita mitkä topiikkien eroja.

Kaksi osallistujaa mainitsivat käyneensä läpi aiheita systemaattisesti hakemalla aihemäärän kautta väliä, jossa aihemallinnuksen tulokset olivat selkeimpiä. Osallistuja esimerkiksi kuvasi tarkastaneensa ensin molemmat ääripäät (20, 39) ja puolivälin (30). Hän kommentoi, että

> 30 aiheen mallinnus puolestaan oli jo melko selkeä verrattuna 20 aiheen mallinnukseen, jossa oli melko monta puurotopicia

Tämän perusteella osallistuja päätti vielä tarkastaa näistä puolivälin (25) ja totesi

> kokenut siinä hukkuvan mitään olennaista, jäin siihen.

Samaan aihemäärään (25) päätyneet osallistujat perustelivat aihemäärää samalla tavoin: aiheet olivat tarpeeksi erillisiä ja erottivat aineistoa tarpeeksi, mutta niissä ei ollut päällekkäisiä aiheita. Toisaalta samankaltaisella perustelulla myös päädyttiin 28 aiheeseen. Vaikka alustavasti tulokset voisi tulkita positiivisiksi: osa osallistujista päätyi subjektiivisella arvioinnilla samaan tulokseen, tulokset myös osoittavat eroja strategian käytössä. Tämä tuo esille, ettei saman valintastrategiankaan käyttö välttämättä takaa samoja tuloksia subjektiivisessa tulkinnassa. Aihemäärän valintaa subjektiivisesti ei siis voida pitää kovinkaan reliabeelina lähestymistapana. Mahdollisesti aihemallinuksen tulkinnan avaaminen ja läpinäkyvyys – esimerkiksi ottaen mallia laadullisesta tutkimuksesta – voisi selkeyttää tätä tilannetta, mutta

aihemallinnusta soveltavissa töissä näin tehdään valitettavan harvoin. Työn toinen osatutkimus osoittaa, että aihemäärän valinta vaikuttaa analyysin erottelukykyyn.

Tuskin on yllättävää, että käyttäjäkokeisiin osallistujien ehdottamat lukumäärät eroavat eivät vain keskenään, vaan myös aineistoista laskettavista tilastollisista mittareista. Kuten kuvasin yllä, myös tilastollisissa mittareissa on eroja ja suositeltu tilastollinen mittari vaihtelee kirjallisuudessa. Paras aihemäärä on harmonisen loglikelihood-arvon perusteella 33 ja perpleksiteetti-arvon mukaan 70. Kuten kuvat 4a sekä 4b näyttävät, loglikelihood-arvo paranee merkittävästi 25 aiheeseen asti ja alkaa huononemaan 50 aiheen jälkeen. Toisaalta perplexiteetti (kuvat 4c sekä 4d) paranee noin 50 aiheeseen asti, jonka jälkeen sen muutokset tasaantuvat.

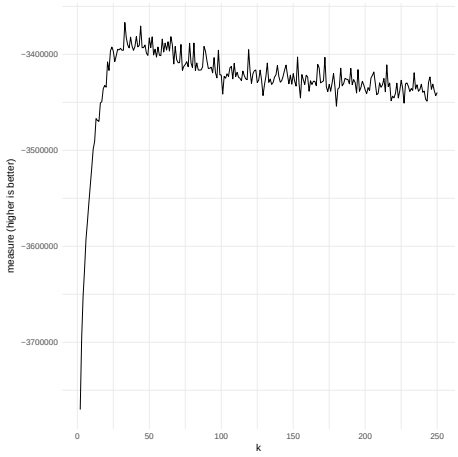
4.3 Keskustelu

Pienimuotoinen käyttäjäkoe osoittaa, ettei yhteiskuntatieteissä käytetty, sisällöllistä tulkintaa korostava, menetelmä valita aiheita ole erityisen luotettava menetelmä. Aihemäärät eroavat tutkijoiden välillä, mikä ohjaamattoman koneoppimisen tapauksessa johtaa erilaisiin tuloksiin (käsittelemme tätä tarkemmin seuraavassa osatutkimuksessa). Myös strategiat aiheiden määrän valintaan ovat erilaisia. Toiset korostivat ennakkokäsitysten, tässä tapauksessa puolueiden käyttöä tunnuspiirteinä ja toiset taas tulkitsivat aiheiden hyvyyttä mallia erottavien tekijöiden kautta. Molemmat lähestymistavat ovat perusteltuja, eikä siis ole ilmeistä, onko mikään käyttäjäkokeeseen osallistuneiden ihmisten suosittelema lukumäärä "paras" kuvaamaan aineistoa.

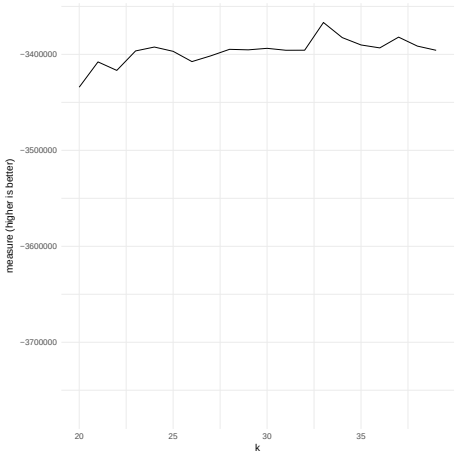
Toisaalta, myös erilaiset tilastolliset mittarit tuottivat erilaisia suosituksia aihemäärästä, kuten kuvassa 4 havaittiin. Tilastollisten mittarien osalta käydään myös aktiivista keskustelua tietojenkäsittelytieteessä eri aiheenvalinnan menetelmien laadusta. Uusin kirjallisuus on päätynyt suosittelemaan harmonista loglikelihood-arvoa, mutta ei ole ilmeistä, että sekään välttämättä tuottaa parhainta tulosta (esimerkiksi, Griffiths & Steyvers 2004, Wallach et al. 2009b). Samalla tavoin kuin subjektiivisen tulkinnan malleissa voidaan siis kysyä, että onko näilläkään mittareilla mahdollista selvittää "parasta" aihemäärää kuvaamaan aineistoa.

Tulokset osoittavat, että eri lähestymistapa voi johtaa hyvinkin erilaisiin tuloksiin. Tässä tapauksessa olemme löytäneet viisi eri vaihtoehtoa parhaasta aihemäärästä. Jokainen niistä on perusteltavissa ja uskoisin niiden olevan hyväksyttyjä myös niistä kirjoitettavissa tieteellisessä artikkelissa[9]. Siksi olemmekin haastavan dilemman edessä: mikä näistä tulisi valita jatkoanalyyseihin parhaiten sopivana? Aihemallinnuksen lisäksi vastaava haaste on olemassa kaikissa ohjaamattoman koneoppimisen menetelmissä: niissä tutkijan tulee aina tehdä jotain rajauksia ja valintoja. Tämä on laskennallisen data-analyysin iso haaste. Kuten Watts (2011) argumentoi, ihmiset ovat erittäin hyviä muodostamaan mielekkäitä tulkintoja *kaikista* tuloksista. Tällöin erilaiset aihemäärät voivat vaikuttaa selkeiltä ja järkeviltä, vaikka aineiston "todellinen" aihejakauma olisi täysin toisenlaisen.

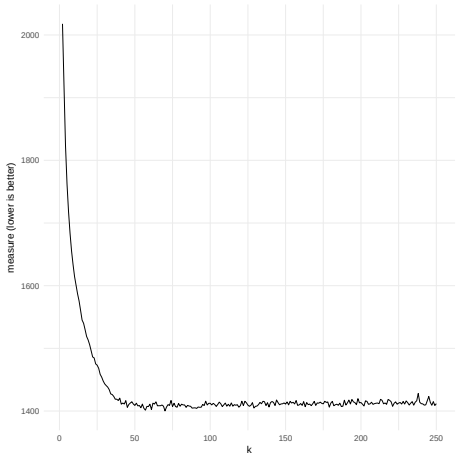
[9] Ehkäpä, mukaillen Dragicevic et al. (2014) artikkelia, missä näytetään p-arvon heikkous toistamalla samaa koeasetelmaa ja näyttämällä erilaisia tuloksia, voitaisiin tästä kirjoittaa viisi eri tieteellistä artikkelia ja jokaisessa keskustella juuri tämän mallin tuottamista erityispiirteistä.



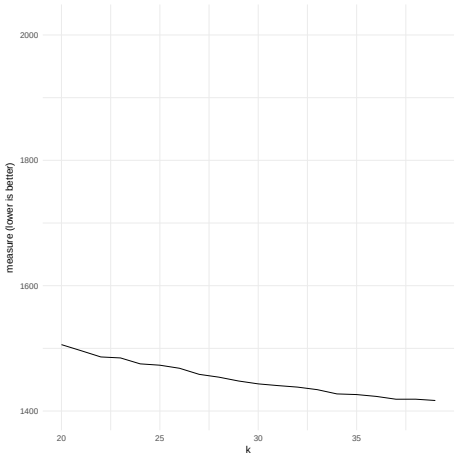
(a) Harmoninen loglikelihood - tarkastelu kn arvoilla 2–250



(b) Harmoninen loglikelihood - tarkastelu kn arvoilla 20–30



(c) Perpleksiteetti - tarkastelu kn arvoilla 2–250



(d) Perpleksiteetti - tarkastelu kn arvoilla 20–30

Kuva 4: Aiheiden määrän sopivuus eri laskennallisilla mittareilla.

Tutkimusmenetelmän toistettavuuden kannalta tämä on erittäin haastavaa, ja siksi subjektiivisten arvioiden sijaan olisikin syytä käyttää tilastollisia mittareita, jotka tuottavat toisinnettavan tuloksen, vaikka sen tulkinnallisuus voikin olla haastavampaa. Siksi ohjaamattomia menetelmiä soveltaville yhteiskuntatietelijöille suositellaan:

Suositus 1 Käytä laskennallisia mittareita mallin parametrien (kuten aiheiden määrän) valinnassa. Kuten jo yllä huomioitiin, tähän on tarjolla useita erilaisia mittareita (Wallach et al. 2009b). Sopivan mittarin valinta vaatii siis alan kehityksen seuraamista. Tätä tekstiä kirjoittaessa monet ovat suositelleet harmooniseen keskiarvoon perustuvaa loglikelihood-mittaria.

Tehtyä pienimuotoista käyttäjäkoetta voidaan toki kritisoida monilla tavoin. Selkein kritiikki on pieni osallistujamäärä, minkä takia ei ole mahdollista tehdä laajempia tilastollisia yleistyksiä aihemäärästä. Toisaalta, koehenkilöillä ei ollut käytössä laajempaa tai yhtenäisempää teoriataustaa mihin tuloksen sidottaisiin: tutkimuskysymyksenä suomalaisen puoluekentän muutosten ymmärtäminen 1900-luvulta tähän päivään on erittäin avoin. Tällöin valitut aihemäärät voivat myös kuvastaa eroavaisuuksia siinä, minkälaiseen kirjallisuuteen ja taustaan koehenkilöt sijoittivat tehtävänannon. Samoin muut taustatekijät ja esimerkiksi mielikuvat Suomen puolueohjelmien sisällöistä voivat hyvin ohjata tässä vaiheessa tapahtuvaa tulkintaa. Toisaalta, jos tutkimuskysymys olisi ollut erilainen, olisiko se johtanut erilaisiin strategioiden käyttöihin? Jos menetelmät ja tulokset eroavat merkittävästi, niin mitä tämä kertoo ohjaamattoman menetelmän mahdollisuuksista tuottaa toistettavia tuloksia? Eräs mahdollisuus olisikin käyttää ohjaamattomia menetelmiä vain aineiston ymmärtämiseen, mutta käyttää ohjatun koneoppimisen menetelmiä yhdistäessä aineistoa yhteiskuntatieteellisiin käsitteisiin, eli opettaa tietokonetta tunnistamaan esimerkkien pohjalta nämä käsitteet (vrt. Nelson 2017).

5 Osatutkimus 2: Aihemäärän vaikutus empiirisiin löydöksiin

Osatutkimuksessa 1 osoitettiin, että aihemäärän valinta on kaikkea muuta kuin jo ratkaistu ongelma. Osatutkimus 2 tarkastelee eri aihemäärillä luotujen aihemallien eroja mahdollisessa jatkoanalyysissä. Osatutkimuksen 1 tutkimuskysymystä mukaillen tässä osatutkimuksessa tavoitteena on kuvailla suomalaisen puoluejärjestelmän muutosta 1900-luvulta tähän päivään. Aiheesta on luontaisesti kirjoitettu erittäin runsaasti, ja seuraava lyhyt katsaus kirjallisuuteen ei tee kunniaa olemassa olevalle tutkimukselle. Tämän lyhyen katsauksen avulla voidaan kuitenkin erottaa muutama selkeä havainto ja oletus puoluekentän muutoksesta: niiden avulla voidaan tarkastella eri aihemäärien vaikusta tutkimuksen tuloksiin.

5.1 Aikaisempi tutkimus puoluejärjestelmien kehittymisestä

Puoluekentän muutosta on pyritty selittämään puolueiden toimintalogiikan muutosten kautta. Farrell & Webb (2000) kuvaavat kuinka poliittiset puolueet ovat muuttuneet

hajautuneista liikkeistä keskitetysti hallittuihin organisaatioihin. Erityisesti laaja muutos on tapahtunut kollektiivi-identiteettien sijaan mahdollisimman laajoiksi yleispuolueiksi tai "catch all-puolueiksi" (esimerkiksi Scarrow 2000). Sellaisenaan yleispuolueen määritelmää on pidetty haastavana (esimerkiksi Maas 2001, Krouwel 2003). Kuitenkin eräs keskeinen piirre määritelmässä on ideologisen pohjan laajentuminen sekä keskittyminen useampiin politiikan alueisiin (esimerkiksi Maas 2001, Krouwel 2003). Yleisesti on argumentoitu, että yllä kuvatut muutokset intressipuolueista yleispuolueiksi ovat ilmeisiä kaikissa länsimaalaisissa demokratioissa.

Yllä oleva kirjallisuus painottui kansainväliseen tilanteeseen. Samankaltaisia tuloksia on tuotu esille myös keskittyen vain suomalaisiin puolueisiin. Esimerkiksi hiljattain julkaistu Mickelsson (2015) jäsentää puoluejärjestelmän kehitystä kuudessa ajanjaksossa. Ensimmäinen ajanjakso (1905-1922) keskittyi kansakunnan sekä valtion järjestäytymiseen. Poliittiset aiheet keskittyivät mm. kielikysymyksen, työväen sekä maaseudun aseman ympärille. Toinen ajanjakso (1923-1939) korosti työväen ja porvareiden välistä eroa sisällissodan seurauksena. Intressiryhmät nousivat edelleen esille kolmantena ajanjakso (1940-1965), jota Mickelsson (2015) kutsuu myös taistelevien intressipuolueiden Suomeksi. Puolueet ajoivat tarkemmin omien intressiryhmiensä etuja, tosin ne kehittyivät kohti yleispuolueita. Intressiryhmien sijaan ideologiat korostuivat neljännellä jaksolla (1966-1978), jota Mickelsson (2015) kuvaa myös rajujen muutosten ajaksi puoluekentällä. Muutokset kumpusivat nuorisoryhmien kautta, esimerkiksi ihmisoikeustyön ja pasifismin kautta. Viides ajanjakso (1979-2007) korosti puolueiden toiminnan muutokseen media- ja markkinapuolueiksi. Modernisaatio on nähtävissä esimerkiksi ekologisen puoluekentän kehittymisenä. Politiikan aiheissa oli murros 'uuteen politiikkaan', elämäntapoihin, feminismiin sekä globaaliin tasa-arvoon. Murros jatkui kuudennella ajanjaksolla (2008-2015) kun vihreiden uuden politiikan näkökannoille muodostui poliittinen vastavoima perussuomalaisista. Myös Arter (1999) on tarkastellut asian tilaa Suomessa. Hän analysoi Keskustan muuttumista yleispuolueeksi ja sen historiallista etenemistä. Hän näki keskeisenä muutoksena keskustan pyrkimyksen laajentaa ideologista pohjaansa vuoden 1962 puolueohjelmassa. Lisäksi puolueen kannattajakunta monipuolistui neljännen ajanjakson aikana (1966–1978). Tämä kuvaa, että Mickelsson (2015) esittämät ajatukset puolueiden muutoksesta tuona aikana saavat tukea myös muusta tutkimuksesta ja onkin ilmeistä, että myös suomalaisessa puoluekentässä on catch all-puolueita.

Yllä olevan kirjallisuuden perusteella on ilmeistä, että myös Suomen puoluekentässä on havaittu jo siirtyminen intressipuolueista yleispuolueiksi 1900-luvun aikana, erityisesti vuosina 1966–1978. Käytämme tätä historiallisen kehityksen muutosta arvioimaan aihemallien luomia tuloksia ja niiden tarkkuutta. Tarkastelemme tätä kysymystä tarkemmin puolueohjelmien aihemallinnuksen valossa seuraavaksi.

5.2 Analyysi ja tulokset

Ensimmäinen vaihe aihemalleihin perustuvissa analyyseissä on antaa aiheille joku merkitys, eli tulkita aihemallien aiheet mielekkäästi. Tulkinnassa voidaan havainnoida aiheissa yleisiä sanoja, tai sanalistoja. Lisäksi aiheista voidaan valita dokumentteja, jotka ovat edustavia. Vain sanalistojen tulkinnan on osoitettu

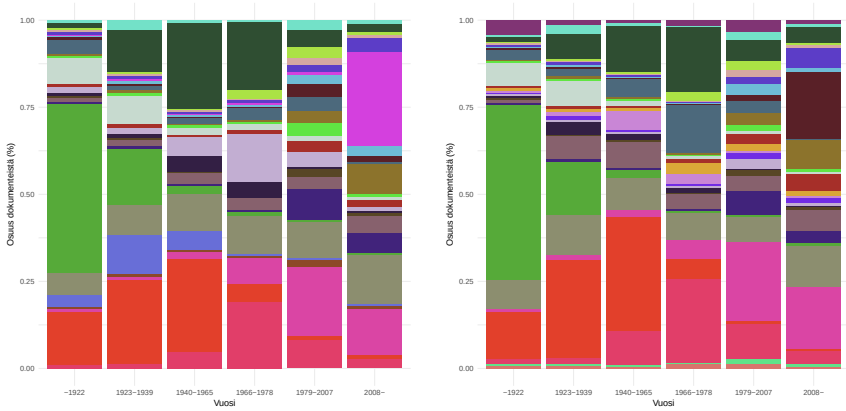
tuottavan erittäin hyviä aihetulkintoja (Aletras et al. 2017). Tällöin edustavien dokumenttien valintaa voidaan käyttää validointivaiheessa. Taulukossa L1 esitetään tulkinnat eri aiheiden määrillä tiivistetyssä muodossa. Tulkinnat perustuvat vain sanalistojen käyttöön, eli aihemallin tuloksia ei ole tätä tarkemmin validoitu.

Aihemallinnuksessa voidaan lisäksi tehdä aiheiden jalostamista, yhteiskuntatieteissä usein teoreettisen käsitteen kautta. Esimerkiksi aiheita voitaisiin ryhmitellä mielekkäiksi kokonaisuuksiksi (vertaa esimerkiksi Strauss & Corbin 1990). Tässä työssä tutkimuskysymyksen kannalta tämä ei ole mielekästä: aihemallien ryhmittymät kuvaavat jo puolueita sekä politiikka-aiheita. Toisaalta, vaikka samalle puolueelle aihemallinnuksen kautta voidaan tunnistaa useampia aiheita (esimerkiksi kokoomus taulukossa L1), ovat nämä sisällöllisesti varsin erilaisia eikä siinä mielessä niiden yhdistäminen ole mielekästä; tutkimuskysymyksenä kun on tarkastella puolueohjelmien muutosta sekä aihemäärien vaikutusta tähän ongelmaan. Toisenlaisella tutkimuskysymyksellä kuitenkin yhdistäminen voi olla tarpeellista ja mielekästä.

Ensimmäinen viite aihemäärän valinnan merkityksestä on nähtävissä taulukosta L1: eri aihemäärät johtavat eri tulkintoihin puolueohjelmien sisällöstä ja niiden aiheista. Jotkin aiheet ovat vakaita ja esiintyvät kaikissa aihemäärissä. Toiset aiheet taas tulevat esille, kun aihemäärää kasvatetaan ja mallin "erottelukyky" kasvaa.

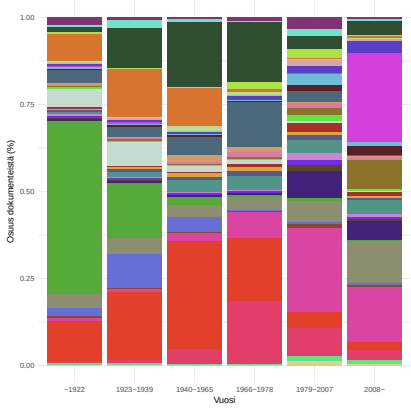
Aihemallinnus auttaa luomaan kokonaiskuvan Suomen puoluekentän muutoksista: kuva 5 näyttää tunnistettujen aiheiden esiintymisen puolueohjelmissa eri tarkasteltavina aihemäärillä käyttäen Mickelsson (2015) esittämää kuuden aikakauden mallia. Kuten kuvista näkyy, pääsääntöisesti aihemallien ryhmittelyt ja antamamme tulkinnat tuottavat samankaltaisen kuvauksen Suomen poliittisesta historiasta. Samaan aikaan eri mallien kautta muodostuva kuva puolueiden historiasta sisältää ilmeisiä eroja. Esimerkiksi ajanjaksolle -1922 25 ja 28 aiheen mallit (kuvat 5a sekä 5b) eivät tuo esille maalaisliiton muutosta keskustaksi, kun taas 33 ja 34 aiheen malleissa (kuvat 5d sekä 5c) aihe näkyy selvemmin (värit ■ ja ■ kuvioissa). Samoin, poliittisten puolueiden ja kentän murrokset (ajanjaksot: 1979-2007 sekä 2008-) ovat selkeämmin esillä 33 ja 34 aiheen malleissa. Kuten kuva 5d näyttää, murrosajanjaksolle ilmeistä on useiden eri poliittisten aiheiden esiintyminen, jotka ovat olleet marginaalissa aikaisempina vuosina. Tämä muutos on vähemmän ilmeinen 25 aiheen mallissa (kuva 5a), jossa väriskaalan muutos on vähäisempi. Vastaavia pienehköjä eroja aiheiden esiintymisessä on nähtävissä enemmän. Tämä antaa lisää osviittaa aiheiden määrän merkityksestä aihemallinnusprosessissa. Kuten näytimme, eri aihemäärät voivat johtaa erilaiseen tulkintaan taustalla olevasta ilmiöstä, tässä tapauksessa poliittisesta historiasta. Palaamme tähän tarkemmin tämän osatutkimuksen keskustelussa.

Mutta onko puolueohjelmien tarkastelun avulla mahdollista nähdä muutos intressipuolueista yleispuolueiksi? Ensimmäinen vaihe kysymykseen vastaamisessa on luonnollisesti *operationalisoida* käsitteinä toisaalta intressipuolue ja toisaalta yleispuolue. Yleispuolueelle ja intressipuolueelle on tarjottu useita erilaisia kriteereitä: yleispuoluetta kuvaa niin niiden sisäinen organisoituminen kuin suhde äänestäjäkuntaan (Kirchheimer 1990, Maas 2001, Krouwel 2003). Puolueohjelmien analyysissä eräs keskeinen määritelmä on, että intressipuolueilla on selvä ja pieni joukko aiheita, joissa puolue on aktiivinen kun taas yleispuolueiden puolueohjelma

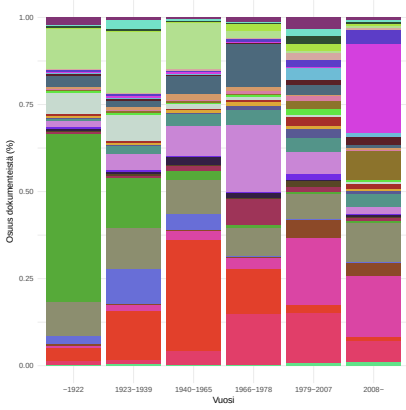


(a) 25 aiheen malli

(b) 28 aiheen malli



(c) 34 aiheen malli



(d) 33 aiheen malli

Kuva 5: Aiheiden jakautuminen puolueohjelmissa. Aiheet väritetty tulkintojen mukaan yhteneväisesti kuvaiden välillä. Tulkinnat esitetty taulukossa L1.

käsittelee useampia aiheita ja ottaa laajemmin kantaa yhteiskuntaan. Tämä on mahdollista edelleen operationalisoida aihemallinnuksen kautta: yleispuolueen puolueohjelma ottaa enemmän ja selkeämmin kantaa useaan aiheeseen, kun taas intressipuolue keskittyy tarkemmin omalle kannattajakunnalle keskeisiin teemoihin.

Kirjallisuuden mukaista kehitystä on tapahtunut: aiheiden määrän kasvu näkyy kuvassa 5 värien määrän kasvuna kaikissa malleissa vuosien -1922 ja 1978 välillä. Visuaalisen tarkastelun lisäksi ilmiötä voidaan tarkastella aihejakauman muutoksien kautta. Vaikka yksittäinen aihe ei *aina* kuvaa tiettyä intressiryhmän etujen ajamista – keskeinen piirre yleispuolueen ja intressipuolueen välillä – taulukko L1 näyttää, että monet aiheet on nähtävissä myös tätä kautta. Poikkeuksen tähän muodostaa kunkin puolueen erityiset puolueaiheet, joissa jokainen puolue on kuitenkin varsin hyvin edustettuna. Koska aiheilla on myös intressiryhmiin liitettäviä merkityksiä, aihemallinnuksen aiheita voi käyttää arvioimaan puolueiden halua viestittää olevansa intressipuolueita.

Laskemme siis kunkin puolueen osalta, kuinka monta aihetta niillä on merkittäviä. Merkittävyyden kriteerinä olemme käyttäneet 10% rajaa: jos puolueohjelma mainitsee tietyn aiheen yli 10% arvosta, niin puolueohjelma tuo esille tämän aiheen merkittävänä[10] Tämän kaltainen rajaus on tarpeen, koska aihemallinnuksessa jokainen aihe on jollain osuudella osana jokaista puolueohjelmaa. Tämän jälkeen vertaamme kaikkia puolueohjelmia keskenään sen perusteella, kuinka monta teemaa kullakin puolueella on esillä. Haasteen analyysiin toki tuo puolueiden sisäisen kehittyminen: onko mielekästä verrata 1980-luvun liikkeitä samalla tavalla kuin jo vakiintuneita toimijoita? Mickelsson (2015) mukaan kehitys intressipuolueista laajemmiksi ideologiapuolueksi tapahtui vuosina 1966–1978. Lisäksi, 1980-luvulla on samaan aikaan ollut olemassa sekä laajoja yleispuolueita että yhden asian liikkeitä, kuten ympäristöryhmittymiä, mikä voisi haastaa analyysiä[11]. Tämän takia rajaudutaan tässä analyysissä keskitytään arvioimaan muutosta ennen vuotta 1966 ja vuosien 1966 ja 1978 välissä merkittävien aiheiden määrän kannalta.

Taulukosta 3 nähdään, että puolueohjelmien merkittävien aiheiden määrä on kasvanut jälkimmäisellä tarkastelujaksolla. Tämä tukee ajatusta siirtymisestä kohti yleispuolueita, ei rajattuihin teemoihin keskittyviä intressipuolueita. Taulukko myös osoittaa aihemallinnuksen aihemäärän merkityksen tuloksen kannalta. 25, 33 ja 34 aiheen malleilla ero on tilastollisesti merkittävä Wilcoxonin mittarilla, kun taas 28 aiheen mallilla tilastollista eroa ei ole.[12] Tilastollisesti merkitsevät

[10] Kymmenen prosentin raja-arvo on varsin mielivaltainen: jollekin toiselle merkittävyyden raja voisi olla 5%, 7.5% tai 12.5%. Ajatuksena taustalla on, että jos puolueohjelma ottaa kantaa johonkin aiheeseen vähintään tai yli kymmenyksellä koko tekstiaineistosta, on tämä aihe puolueohjelmassa merkittävästi esillä.

[11] Eräs ratkaisu olisikin tarkastella liikkeiden ja puolueiden institutionalisoitumisprosessia, eli arvioida kuinka kunkin puolueen merkittävien aiheiden määrä on muuttunut vuosien varrella. Kuitenkin tässäkin lähestymistavassa edessä olisi ollut ilmeisiä haasteita: puolueiden välillä voi olla eroja niiden kehitysnopeudessa kohti yleispuolueita, toisaalta puolue itsessään on jo institutionalisoitunut ja analyysin olisi vaikea huomioida muita puolueiden toimintaan vaikuttavia tekijöitä – kuten teknologian kehitystä (vrt. Farrell & Webb 2000). Tämän takia on argumentaation selkeyden kannalta selkeämpää rajautua ajanjaksoon ennen uusien liikkeiden syntymistä ja arvioida puolueohjelmia tämän aikakauden kautta.

[12] Wilcoxonin testi on parametriton testi yhteiskuntatieteilijöille tutummasta t-testistä, jota

Taulukko 3: Puolueohjelmassa olevien merkittävien aiheiden määrän keskiarvo ennen vuotta 1965 ja 1966–1978.

Malli	–1965	1966–1978		
25 aihetta	2.41	3.34	$p = < 0.01$	$(W = 413)$
28 aihetta	2.44	2.86	$p = 0.12$	$(W = 622.5)$
33 aihetta	2.64	3.41	$p = 0.006$	$(W = 516)$
34 aihetta	2.71	3.28	$p = 0.03$	$(W = 567.6)$

mallit myös esittelevät ilmiötä osittain eri tavoin: 25 aiheen mallilla ero on huima: vuosina 1966–1978 puolueet käsittelivät yhtä aihetta enemmän kuin aikaisemmin, kun taas 33 ja 34 aiheen malleilla ero on noin puolen aiheen verran. Taulukon 3 osittain ristiriitaiset tulokset tuovat erinomaisesti esille aihemallinnuksen haasteen käytännön tutkimuksessa. Aihemallinnuksen aikana tehnyt valinnat tuovat hyvinkin erilaisia näkökulmia Suomen puoluekentän muutokseen ja pahimmillaan johtivat (28 aiheen tapauksessa) tulokseen, ettei muutosta intressipuolueesta yleispuolueeksi ole havaittavissa tällä tavoin operationalisoituna. Kuitenkin, toiset aihemallinnuksen tulokset tunnistavat puolueiden kehittymisen: puolueohjelmat ottavat merkittävästi kantaa useampaan aiheeseen vuosina 1966–1978 kuin aikaisempina vuosina, eli puolueet ovat muuttuneet yleispuoluemaisiksi toimijoiksi.

5.3 Keskustelu

Tärkein osatutkimuksen havainto on, että aiheiden jakaumien kautta esille tuodut muutokset ovat mahdollista sijoittaa Mickelsson (2015) esittämään jakoon Suomen puolueiden kehityksestä. Aihemallinnuksen kautta on selkeästi nähtävissä esimerkiksi muutos maaseutuyhteiskunnasta teolliseen yhteiskuntaan. Samoin myös muutos intressipuolueesta yleispuolueeksi on mahdollista operationalisoida ja tätä kautta mitata aihemallinnuksen kautta.

Koska yllä olevat muutokset olivat laajasti tiedossa kirjallisuudessa, luvun varsinainen anti onkin enemmän menetelmällinen. Jatkaen osatutkimuksen 1 kysymyksen asettelua, miten aihemäärä vaikuttaa analyysin tuloksiin? Niin kuva 5 kuin taulukko 3 näyttävät, että aihemäärän valinta on merkittävä osa tutkimusprosessia ja vaikuttaa tuloksiin. Nykyisillä tavoilla tehdä aihemallinnusta voitaisiin puolustaa jokaista tässä analyysivaiheessa käytettyä aihemäärää: 25, 28, 34 tai 33 aihetta. Tämä johtaa varsin haastavaan kysymykseen: mikä kuvista 5a, 5b, 5c tai 5d on "oikea" ilmentymä puolueohjelmista ja niiden historiallisesta kehityksestä. Välttääksemme tätä ongelmaa, esittelimme jo edellä suosituksen 1 suosia tilastollista mittaria aihemäärän valinnassa. Näin tekemällä vältetään keskustelu esimerkiksi samanlaisella strategialla valitun 25 ja 28 aiheen välillä. Molemmat olivat yhtä hyvin perusteltuja ja samankaltaisen valintastrategian muodostamia, mutta joista vain toinen toi esille tilastollisesti merkittävän eron puolueohjelmien aiheiden määrän analyysissä taulukossa 3. Tätä kautta kysymys

käytän koska ei ole syytä olettaa aineiston olevan normaalijakautunut. Koska tässä on tehty useita p-testauksia, taulukossa esitettävät luvut on korjattu Bernoulli-kertoimella.

muutoksesta intressipuolueesta yleispuolueeksi saa erilaisia vastauksia.

Subjektiivista ja tulkinnallisuutta lähestymistapana puolustavat henkilöt voisivat argumentoida, että eri aihemäärien avulla laskettujen aihemallien tuottamat erot ovat vähäisiä ja siksi tulkittavuuden korostaminen on hyväksyttyä. Tämä argumentti ei ole ilmiselvästi virheellinen. Esimerkiksi kuva 5 tuottaa samanlaisia trendejä ilmiöistä ja laajat, yhteiskuntatason merkittävät muutokset – kuten puolueiden aihemäärän laajentuminen vuoteen 1979 asti sekä uusien poliittisten liikkeiden nousu 1979–2007 – ovat esillä kaikissa kuvissa. Kuitenkin taulukko 3 tuo tarkemmin esille lähestymistavan ongelman: löydösten tilastollinen merkitsevyys voi muuttua aihemäärän valinnan seurauksena.

Aihemäärän valintaan liittyvät ongelmat näyttävät, että aihemallinnuksen käytön empiirisessä analyysissä pitäisi olla huolellista ja tarkkaa. Löydös sellaisenaan jatkaa käynnissä olevaa varsin vilkasta keskustelua aihemallinnuksen eri vaiheiden, kuten esiprosessoinnin (Schofield & Mimno 2016, Denny & Spirling 2018) sekä mallinnuksen parametrien valinnan vaikutuksesta (Wallach et al. 2009a) aihemallinnuksen lopputuloksiin. Tässä työssä esitellyt löydökset nostavat esille jälleen yhden uuden haasteen aihemallinnuksen soveltamiseen empiirisessä tutkimuksessa. Osatutkimus 2 osoittaa, että aihemäärän valinnalla on merkitystä tutkimuksen kannalta.

Toisaalta, osatutkimus 2 myös tuo esille tulkinnan ja luokittelun haasteet aihemallinnuksessa. Esimerkiksi tässä työssä on päätetty esittää kukin aihe erillisenä kokonaisuutena. Kuten jo yllä keskusteltiin, olisi myös mahdollista luokitella aiheita laajempiin yläkategorioihin ihmisen tekemän tulkinnan mukaan. Esimerkiksi aiheet “kapitalismi ja sosialismi” sekä “sosialismi” voitaisiin myös yhdistää saman yläotsikon (‘metakoodin’) alle, jolloin yhdistelmäaihe mahdollisesti kuvaisi tarkemmin ristiriitaulottuvuutta. Samoin kansainvälisyyden ja globalisaation aiheet voisi olla mielekästä yhdistää. Tällä tavalla tutkijan tulkinnat tulevat jälleen osaksi aineistoa, mutta voivat selkeyttää aihemallinnuksen tulosten tulkintaa, jos aihemäärä on erityisen suuri.

Avoimena kysymyksenä nouseekin siis, miten aihemallinnuksen soveltajaa voitaisiin paremmin tukea tässä analyyttisessä vaiheessa. Tässä artikkelissa pyrittiin aiheiden valintoja perustelemaan sekä olemassa olevan aihepiiriä käsittelevän kirjallisuuden (Mickelsson 2015) sekä yleisen politiikan tutkimuksen teorian kannalta (Kirchheimer 1990, Maas 2001, Krouwel 2003). Aihepiirin kirjallisuutta käytettiin hyödyksi sekä aihemallinnuksen historiallisten muutosten tulkinnassa että myös yksittäisten aiheiden valinnassa – esimerkiksi Luonnonlain puolueen tunnistaminen perustui tähän kirjallisuuteen tutustumiseen. Toisaalta, laajempi teoriakatsanto mahdollisti aihemallinnuksen käyttämisen ei lopullisena tuloksena vaan välivaiheena, jota jatko-operationalisoitiin ja käytettiin teorian tilastollisessa tarkastelussa. Tällöin tehdyille valinnoille voidaan hakea tukea jo olemassa olevasta kirjallisuudesta.

Toisaalta, aina ei ole olemassa yhtä vahvaa kirjallisuutta, jonka päälle analyysin voisi rakentaa. Tämän takia viimeaikaisessa kirjallisuudessa on laajasti pohdittu miten laskennallinen analyysi voisi tukeutua ja täydentää perinteistä laadullista työtä triangulaation hengessä (Laaksonen et al. 2017, Nelson 2017, Muller et al. 2016). Kolmas vaihtoehto olisi nähdä aihemallinnus (sekä vastaavat ohjaamattoman koneoppimisen menetelmät) eräänlaisena grounded theory-vaiheena. Tämä onkin mielekästä jos aiheesta ei ole etukäteen tiedossa mahdollisia luokkia ja niiden

tulkintoja, vaan kyseessä on avoimempi tutkimussuunta. Suomenkielisinä esimerkkeinä tämän kaltaisesta työstä voi nähdä viimeaikaiset julkaisut agendan hallinnasta sekä kehyksistä (Laaksonen & Nelimarkka 2018, Ylä-Anttila et al. 2018). Kuitenkaan näissä töissä ei noudateta perinteisen grounded theory-menetelmän henkeä esimerkiksi uudelleenkoodauksen tai muistiinpanojen ('memoing') osalta (vertaa Strauss & Corbin 1990). Ohjaamattomat tekstianalyysimenetelmät ovat valitettavasti vielä toistaiseksi avoimia, eikä yleisesti hyväksyttyjä hyviä käytänteitä aiheiden tunnistamiseen, nimeämiseen ja ryhmittelyyn ole olemassa. Tämän takia tähän vaiheeseen on toistaiseksi tarpeen kiinnittää erityistä huomiota. Yllä on esitelty kolme erilaista näkökulmaa aiheiden tulkintaan ja sen soveltamiseen, jotka voidaan tiivistää seuraavasti:

Suositus 2 Aiheiden tulkintaan ja niiden käyttöön voidaan käyttää jotain seuraavista kolmesta lähestymistavasta:

- käyttäen olemassa olevaa kirjallisuutta joko sisällöllisenä apuvälineenä tai käyttäen aihemallinnusta välineenä olemassa olevan teorian arvioimiseen
- soveltaen aihemallinnusta sekä aineiston muita analyyseja trianguloiden toistensa kanssa rinnakkain
- kuvaten kuinka aihemallinnuksen tuloksissa tehtiin uudelleenryhmittelyjä sekä kirjaamalla omia havaintoja ja tulkintoja systemaattisesti kuten aineistolähtöisessä grounded theory-prosessissa on ohjeistettu.

Näistä ensimmäinen suositus on samankaltainen kuin ensimmäisessä osatutkimuksessa tunnistettua aiheiden määrän valinnan strategiaa. Toisessa kahdesta tunnistetusta stretegiasta määrän valintaan perustui olemassa olevan ennakkokäsityksen hyödyntämiseen mallin valinnassa. Tämän strategian kohdalla onkin syytä olla varovainen, ettei tutkimuksen aikana ensin perustella mallin valintaa olevalla teorialla ja tämän jälkeen perustella mallin mielekkyyttä mallin sopivuudella teoriaan.

6 Keskustelu ja johtopäätökset

Viimeaikoina myös perinteisissä yhteiskuntatieteissä on herännyt mielenkiintoa käyttää ohjaamattomia koneoppimismenetelmiä, kuten aihemallinnusta. Uusien menetelmien avulla tutkimuksen lähestymistavat voivat olla datavetoisia, eksploratiivisia, iteratiivisia sekä suuriin aineistomääriin sopivia (Kitchin 2014). Ilmiselvää kuitenkin on, ettei (ohjaamattomien) koneoppimismenetelmien käytön ei tulisi olla 'teoriatonta' tai 'historiatonta.' Massadatasta ei ilman ymmärrystä kontekstista sekä teorioista voida sanoa mitään (Frické 2015, Boyd & Crawford 2012). Yhteiskuntatieteilijöiden perinteinen integroituminen alansa oppihistoriaan ja kirjallisuuteen voikin mahdollistaa laadukkaan yhteiskuntatieteellisen tutkimuksen myös (ohjaamattomia) koneoppimismenelmiä käyttäessä (Wallach 2018, Grimmer 2015).

Tämä tutkimus osaltaan ilmentää, miten ohjaamattomien koneoppimismenetelmien sovelluksissa voidaan sitoutua alan kirjallisuuteen. Sitoutuminen Suomen puoluejärjestelmän historiaan tuki joidenkin aiheiden nimeämistä, esimerkiksi Luonnonlaki-puolueen

kohdalla, jossa aiheissa oli sanoja kuten ‘tietoisuus’ sekä ‘luonnonlaki’. Osatutkimuksessa 2 ehdotetaankin, että aiheiden tulkinnassa ja niiden arvioimisessa tilastollisten menetelmien, kuten aiheiden sisäisen yhteneväisyyden mittaamisen (Chang et al. 2009, Towne et al. 2016), tukena voi myös käyttää teoriaan tai triangulaatioon perustuvia lähestymistapoja. Osatutkimus 2 perustui sitoutumiseen teoriaan, mutta mahdollisuuksia on myös etnografian sekä muiden laadullisten menetelmien käytössä ja yhdistämisessä massadataan (Laaksonen et al. 2017, Muller et al. 2016).

Puolueohjelmien muutosta sekä massapuolueiden kehittymistä kuvaavan tutkimuksen sijaan artikkelin motivoivina tutkimuskysymyksinä oli ymmärtää aihemallinnuksen metodologisia haasteita yhteiskuntatieteille. Kuten osatutkimus 1 toi esille, yhteiskuntatieteellinen tutkimus on perinteisesti suosinut tulkinnallisuutta korostavia lähestymistapoja, joissa tutkija arvioi mikä aihemäärä on helpoiten tulkittavissa. Osatutkimus 1 tarkasteli tutkijoiden lähestymistapoja määrittää tulkinnallisuuden kannalta paras aihemäärä. Tutkimuksessa havaitaan eroja määrissä: neljä eri osallistujaa ehdottivat parhaimmiksi aihemääriksi 25:tä, 28:aa tai 34:ää. Eroja oli myös tulkittavuuden ymmärtämisessä: toiset tutkijat korostivat eri näkökulmien huomioimista ja mukaan tuomista, kun taas toiset pyrkivät yksinkertaistamaan ja vähentämään aiheiden määrää. Osatutkimus 2 näytti, että aiheiden määrällä on vaikutusta analyysin tuloksiin.

Mitä nämä kaksi osatutkimusta kertovat laajemmin ohjaamattomista koneoppimismenetelmistä ja niiden soveltumisesta yhteiskuntatieteelliseen tutkimukseen? Ensin on syytä ymmärtää minkälaisia lupauksia nämä menetelmät ja laskennallinen yhteiskuntatiede on tarjonnut. Laskennallista yhteiskuntatiedettä esittelevässä artikkelissaan Lazer et al. (2009) kutsuvat laskennallisia menetelmiä uudeksi mikroskoopiksi. Heidän mukaan laskennallisten menetelmien avulla voidaan tarkastella yhteiskuntaa uusilla tavoilla. Mutta, minkälainen mikroskooppi on kyseessä? Kuten Giere (2010) huomauttaa – käyttäen linssejä esimerkkeinä – tieteellinen tieto ei ole koskaan objektiivista, vaan värittynyt käytettyjen instrumenttien kautta. Ohjaamattomat menetelmät ovat haastavia instrumentteja, koska mittavälineen soveltuvuus ongelmaan selviää usein vasta kun analyysi on tehty. Esimerkiksi aihemallinnusprosessin keskeinen vaihe, ohjaamattoman koneoppimisen kautta syntyneiden aiheiden nimeäminen ja tulkinta on mahdollista vasta kun ne on jo tuotettu. Sitä ennen on täysin tuntematonta, minkä kaltaisia ryhmiä aineistosta “nousee.” Nämä menetelmät ovat usein myös satunnaisuutta käyttäviä, jolloin saman menetelmän käyttäminen ei välttämättä johda sellaisenaan samoihin tuloksiin. Symons & Alvarado (2016) argumentoivatkin, että laskennallisen analyysin virhelähteisiin tulisi keskittyä aiempaa enemmän tulosten tulkinnassa.

Tarkoittaako yllä esitetty laskennallisten menetelmien kritiikki sitä, ettei tieteellisessä yhteisössä voida ajatella ohjaamattomien koneoppmismenetelmien tuottavan tieteellistä tietoa? Mielestäni tilanne ei ole näin vakava. Esimerkiksi Giere (2010) kritiikissä huomio keskittyy siihen, että tieteellinen havainnointi on mahdollista kuvata läpinäkyvästi. Tämä pätee myös esimerkiksi laadulliseen tutkimukseen, missä on kuitenkin tarpeen reflektoida myös tutkijan roolia osana tulkintaa ja tuloksia. Tällä hetkellä tämä ei kuitenkaan ole yleinen tapa ohjaamattoman koneoppimisen käyttämisessä (katso loppuviite 8). Kuitenkin, henkilön omalla

taustalla ja kokemuksella on merkitystä esimerkiksi teoriaviitekehysten valinnassa sekä laajemmin tulkinnassa. Esimerkiksi osatutkimuksessa 1 havaitut erot osaltaan heijastelivat eri koehenkilöiden taustojen merkitystä. Tämän takia ohjaamattoman koneoppimisen sovelluksiin voitaisiin pyrkiä soveltamaan samanlaisia käytänteitä kuin laadullisessa tutkimuksessa. Tämän tavoite on nostaa esille myös ohjaamattomassa koneoppimista soveltavien tutkimusprojektien useat erilaiset valinnat.

Suositus 3 Tutkimustyön raportointi tulisi olla refleksiivisempää. Siinä olisi kuvattava tutkijoiden taustoja ja pohtia mahdollisia syitä tulkintoihin. Lisäksi tutkimustyön aikana tehtyjä ei-julkaistuja analyysejä, tehtyjä valintoja sekä rajauksia on tarpeen esitellä.

Samaan aikaan kun laskennallinen yhteiskuntatiede suosittaa algoritmien käyttöä tutkimusongelmien ratkaisemiseen, kriittisen algoritmitutkimuksen koulukunta tuo esille algoritmisten järjestelmien yhteiskunnallisia vaikutuksia, erityisesti näkymätöntä valtaan (esimerkiksi Kitchin 2017, Gillespie 2012). Kriittisen algoritmitutkimuksen sanoma on – tiivistäen – että algoritmien kehittäjillä ja suunnittelijoilla on näkymätöntä valtaa (esimerkiksi Beer 2017). Esimerkiksi algoritmien aiheuttama syrjintä päätöksentekotilanteessa on ollut aktiivinen tutkimusalue (muun muassa Burrell 2016). Hiljaittain van Es et al. (2018) ovat pyrkineet nostamaan samaa ongelmanasettelua esille myös tieteiden piirissä kritisoidessaan tutkimusta varten luotuja työkaluja ja vaatiessaan niiden tarkemaa tutkimusta.

Kritiikki on erittäin osuvaa myös laskennallisen yhteiskuntatieteen piirissä. Onkohan mahdollista, että laskennallisessa tutkimuksessa käytettävissä koneoppimisalgoritmeissä on samanlaisia vallankäyttöön liittyviä piirteitä? Esimerkiksi tekstianalyysin (tai, teknisemmin luonnonollisen kielen käsittelyn) menetelmät, systemaattisiin väärintulkintoihin? Esimerkiksi sanojen irrottaminen lausejärjestyksestä (*bag of words*) voi johtaa systemaattisiin virhetulkintoihin. Toisaalta, onko syytä suosittaa nimenomaisesti tiettyä mittaria aihemäärän valintaan, kuten tässä työssä on tehty?

Yllä esitetyt kysymykset ovat avoimia, mutta tieteellisen tutkimustyön kannalta keskeisiä. Huolimaton tai virheellinen ohjaamattoman menetelmän käyttö saattaa "pakottaa" menetelmän tuottamaan tiettyjä tuloksia – ja nämä tulokset voivat luoda vääristyneen näkökulman maailmaan. Tämän takia osatutkimuksen 2 pohdinnat tulosten tulkinnasta korostivat tarvetta pohtia löydöksiä suhteessa toisaalta muihin menetelmiin ja toisaalta muihin teorioihin. Toistaiseksi kriittisen algoritmitutkimuksen kirjallisuus ei ole käsitellyt mitään tiettyjä menetelmiä tai lähestymistapoja tarkasti tieteellisen tutkimustyön osana. Siksi toistaiseksi voidaan antaa vain seuraava yleinen suositus:

Suositus 4 Laskennallisen yhteiskuntatieteen soveltajien tulee seurata menetelmäkehityskeskustelun lisäksi kriittisen algoritimikirjallisuuden havaintoja esimerkiksi algoritmien puolueellisuudesta.

Artikkeli on laajentanut suomalaista keskustelua mahdollisuuksista tekstiaineistojen analyysiin automaattisesti. Olen artikkelissa kyseenalaistanut yhteiskuntatieteessä yleisiä lähestymistapoja käyttää aihemallinnusta, ja samankaltainen kritiikki

voidaan laajentaa koskemaan muita ohjaamattomia menetelmiä. Keskusteluni tutkimusprosessista sekä sen esimerkinomainen sovellus näyttävätkin, etteivät ohjaamattomat menetelmät ole automatisoitu prosessi, vaan vaatii aina tutkijan omaa tulkintaa ja päätöksentekoa.

Tutkimuksen ensimmäisessä osatutkimuksessa havaittiin, että tulkinnallisuutta korostavassa aihemallinnustyössä aihemäärät vaihtelevat runsaasti. Toisessa osatutkimuksessa taas näytettiin, että aihemäärällä oli vähäisiä, mutta merkillepantavia eroja empiirisissä tuloksissa. Yhdessä osatutkimuksen nostavat esille haasteita aihemallinnuksen luotettavuuteen ja toistettavuuteen liittyviä kysymyksiä. Kysymysten ratkaisemiseksi suositeltiin, että aihemallinnuksessa siirryttäisiin käyttämään tilastollista aiheiden määrän valintaa sekä selkeästi muodostettaisiin yhteys olemassa olevaan yhteiskuntatieteelliseen teoriaan. Lisäksi laskennallisia menetelmiä kohtaan voi olla syytä olla kriittinen sekä vaatia tarkempaa raportointia tutkimusprosessista.

Viitteet

Nikolaos Aletras, Timothy Baldwin, Jey Han Lau, & Mark Stevenson. Evaluating topic representations for exploring document collections. *Journal of the Association for Information Science and Technology*, 68(1):154–167, jan 2017.

David Arter. From class party to catchall party?: The adaptation of the Finnish Agrarian-Center Party. *Scandinavian Political Studies*, 22(2):157–180, 1999.

Deborah Bandalos & Meggen Boehm-Kaufman. *Four common misconceptions in Explanatory Factor Analysis*. Routledge,

David Beer. The social power of algorithms. *Information, Communication & Society*, 20(1):1–13, 2017.

David M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4): 77–84, apr 2012.

David M Blei, Andrew Y Ng, & Michael I Jordan. 10.1162/jmlr.2003.3.4-5.993. *CrossRef Listing of Deleted DOIs*, 1:993–1022, 2000.

Danah Boyd & Kate Crawford. Critical Questions for Big Data. *Information, Communication & Society*, 15(5):662–679, jun 2012.

Jenna Burrell. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society*, 3(1):2053951715622512, 2016.

Jonathan Chang, Sean Gerrish, Chong Wang, & David M Blei. Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems 22*, pages 288—-296, 2009.

Claudio Cioffi-Revilla. Computational social science. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(3):259–271, may 2010.

Matthew James Denny & Arthur Spirling. Text Preprocessing For Unsupervised Learning: Why It Matters, When It Misleads, And What To Do About It. *Political Analysis*, 26(02):168–189, apr 2018.

Pierre Dragicevic, Fanny Chevalier, & Stephane Huot. Running an HCI experiment in multiple parallel universes. In *Proceedings of the extended abstracts of the 32nd annual ACM conference on Human factors in computing systems - CHI EA '14*, pages 607–618, New York, New York, USA, 2014. ACM Press.

Leandre R Fabrigar, Duane T Wegener, Robert C MacCallum, & Erin J Strahan. Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, 4(3):272–299, 1999.

David M Farrell & Paul Webb. *Political parties as campaign organizations*, pages 102–128. Oxford University Press Oxford,

Martin Frické. Big data and its epistemology. *Journal of the Association for Information Science and Technology*, 66(4):651–661, apr 2015.

Ronald N Giere. *Scientific perspectivism*. University of Chicago Press,

Tarleton Gillespie. The relevance of algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, pages 167–194. 2012.

Derek Greene, Derek O'Callaghan, & Pádraig Cunningham. How many topics? Stability analysis for topic models. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8724 LNAI(PART 1):498–513, 2014.

Thomas L Griffiths & Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(Supplement 1):5228–5235, apr 2004.

Justin Grimmer. We Are All Social Scientists Now: How Big Data, Machine Learning, and Causal Inference Work Together. *PS: Political Science & Politics*, 48(01):80–83, jan 2015.

Justin Grimmer & Brandon M. Stewart. Text as Data: The Promise and Pitfalls of Automatic Content Analysis Methods for Political Texts. *Political Analysis*, 21(3):267–297, jul 2013.

Carina Jacobi, Wouter van Atteveldt, & Kasper Welbers. Quantitative analysis of large amounts of journalistic texts using topic modelling. *Digital Journalism*, 4(1):89–106,

Steven J. Jurek & Anthony Scime. Achieving Democratic Leadership: A Data-Mined Prescription. *Social Science Quarterly*, 95(1):97–110, mar 2014.

Otto Kirchheimer. The catch-all party. *The West European Party System*, pages 50–60,

Rob Kitchin. Big Data, new epistemologies and paradigm shifts. *Big Data & Society*, 1(1):1–12, apr 2014.

Rob Kitchin. Thinking critically about and researching algorithms. *Information, Communication & Society*, 20(1):14–29, jan 2017.

André Krouwel. Otto Kirchheimer and the catch-all party. *West European Politics*, 26(2):23–40, 2003.

Salla-Maaria Laaksonen & Matti Nelimarkka. Omat ja muiden aiheet : Laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta. *Politiikka*, 60(2):132–147,

Salla-Maaria Maaria Laaksonen, Matti Nelimarkka, Mari Tuokko, Mari Marttila, Arto Kekkonen, & Mikko Villi. Working the fields of big data: Using big-data-augmented online ethnography to study candidate–candidate interaction at election time. *Journal of Information Technology and Politics*, 14(1):110–131, jan 2017.

David Lazer, Alex Pentland, Lada Adamic, Sinan Aral, Albert-Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, Tony Jebara, Gary King, Michael Macy, Deb Roy, & Marshall Van Alstyne. Social science. Computational social science. *Science (New York, N.Y.)*, 323:721–723, 2009.

Karen E. C. Levy & Michael Franklin. Driving Regulation: Using Topic Models to Examine Political Contention in the U.S. Trucking. *Social Science Computer Review*, 32(2):182–194, apr 2014.

Willem Maas. Catch-all parties. In *Reader's Guide to the Social Sciences*, pages 167—-168. London: Fitzroy Dearborn,

Jari Metsämuuronen. Tutkimuksen tekemisen perusteet ihmistieteissä.

Rauli Mickelsson. *Suomen puolueet: Vapauden ajasta maailmantuskaan.* Vastapaino, Tampere,

John W. Mohr & Petko Bogdanov. Introduction-Topic models: What they are and why they matter. *Poetics*, 41(6):545–569, 2013.

Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, & N. Sadat Shami. Machine Learning and Grounded Theory Method. In *Proceedings of the 19th International Conference on Supporting Group Work - GROUP '16*, pages 3–8, New York, New York, USA, 2016. ACM Press.

Matti Nelimarkka, Salla-Maaria Laaksonen, Mari Marttila, Arto Kekkonen, Mari Tuokko, & Mikko Vili. Influencing the agenda through social media: Online agenda building and normalization during a pre-electoral campaign period.,

Laura K. Nelson. Computational Grounded Theory. *Sociological Methods & Research*, page 004912411772970, nov 2017.

Semi Purhonen & Arho Toikka. "Big datan"haaste ja uudet laskennalliset teksiaineistojen analyysimenetelmät. *Sosiologia*, (1):6–26,

Daniel W. Russell. In Search of Underlying Dimensions: The Use (and Abuse) of Factor Analysis in Personality and Social Psychology Bulletin. *Personality and Social Psychology Bulletin*, 28(12):1629–1646, 2002.

Mike Savage. The 'Social Life of Methods': A Critical Introduction. *Theory, Culture & Society*, 30(4):3–21, 2013.

Susan E Scarrow. Parties without members?: party organization in a changing electoral environment. pages 102–128,

Alexandra Schofield & David Mimno. Comparing Apples to Apple : The Effects of Stemmers on Topic Models. 4:287–300,

Anselm Strauss & Juliet M Corbin. *Basics of qualitative research: Grounded theory procedures and techniques.* Sage Publications, Inc,

John Symons & Ramón Alvarado. Can we trust Big Data? Applying philosophy of science to software. *Big Data & Society*, 3(2):205395171666474, 2016.

W Ben Towne, Carolyn P Rosé, & James Herbsleb. Measuring Similarity Similarly: LDA and Human Perception. *ACM Transactions on Intelligent Systems and Technology ACM Reference Format ACM Trans. Intell. Syst. Technol*, 7(2):25:1–25:29, 2016.

Karin van Es, Maranke Wieringa, & Mirko Tobias Schäfer. Tool Criticism. In *Proceedings of the 2nd International Conference on Web Studies - WS.2 2018*, pages 24–27, New York, New York, USA, 2018. ACM Press.

Hanna Wallach. Computational social science computer science + social data. *Communications of the ACM*, 61(3):42–44, 2018.

Hanna M Wallach, David Mimno, & Andrew Mccallum. Rethinking LDA: Why Priors Matter. *Advances in Neural Information Processing Systems 22*, 22(2): 1973–1981, 2009a.

Hanna M Wallach, Iain Murray, Ruslan Salakhutdinov, & David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, number 4, pages 1–8, New York, New York, USA, 2009b. ACM Press.

Duncan J Watts. *Everything is obvious:* Once you know the answer.* Crown Business,

Tuukka Ylä-Anttila, Veikko Eranti, & Anna Kukkonen. Aihemallinnuksesta kehitysmallinnukseen. *Politiikka*, 60(2):158–156,

Bei Yu, Stefan Kaufmann, & Daniel Diermeier. Classifying Party Affiliation from Political Speech. *Journal of Information Technology & Politics*, 5(1): 33–48, jul 2008.

Liitteet

Liitetaulukko L1: Aihemallin tulkintoja eri aihemäärillä

	Tulkinta	$k = 25$	$k = 28$	$k = 33$	$k = 34$
Suuret kysymykset					
	Kapitalismi ja sosialismi	kapitalismi, kapitalisti, kehitys, kommunisti, puolue, sosialismi, sosialistinen, taistelu, työväenluokka, yhteiskunta	kapitalismi, kapitalisti, kehitys, kommunisti, maailma, politiikka, sosialismi, sosialistinen, yhteiskunnallinen, yhteiskunta	kapitalismi, kapitalisti, kehitys, kommunisti, maailma, politiikka, pääoma, sosialismi, sosialistinen, yhteiskunta	kapitalismi, kapitalisti, kommunisti, kommunistinen, maailma, politiikka, pääoma, sosialistinen, yhteiskunnallinen, yhteiskunta
	Isänmaa	henki, isänmaa, johtaa, kansa, kansallinen, perusajatus, pääoma, suomalainen, suomi, suomia	henki, isänmaa, kansa, kansallinen, perusajatus, pääoma, suomalainen, suomi, suomia, tarkoittaa	henki, isänmaa, kansa, kansallinen, perusajatus, suomalainen, suomi, suomia, tarkoittaa, toimia	henki, isänmaa, kansa, maa, perusajatus, pääoma, suomalainen, suomi, suomia, toimia
	Kansalaisuus	kansa, kansalainen, mahdollisuus, saada, taloudellinen, toiminta, valtio, vapaus, yhteiskunta, yksityinen	kansa, kansalainen, mahdollisuus, sosiaalinen, taloudellinen, toiminta, vapaa, vapaus, yhteiskunta, yksilö	ihminen, kehitys, markkinat, poliittinen, sosiaalinen, tuotanto, vapaus, vasemmisto, yhteiskunta, yritys	kansa, kansalainen, pyrkiä, saada, taloudellinen, tehtävä, toiminta, turvata, vapaus, yhteiskunta
Puolueaiheet					
	Keskusta	henkinen, ihminen, kehitys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomi, turvata	henkinen, ihminen, kehitys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomalainen, suomi	ihminen, ihmisyys, kehitys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomalainen, tulevaisuus	aineellinen, henkinen, ihminen, ihmisyys, keskusta, keskustapuolue, luonnonvara, maakunta, maaseutu, suomalainen
	Maalaisliitto				elämä, kansa, kansallinen, lainsäädäntö, maa, maalaisliitto, maaseutu, olo, sivistys, taloudellinen
	Maalaisliitto			kansa, maa, maalaisliitto, maaseutu, maatalous, olo, sivistys, talous, työ, valtio	alue, elintaso, kuivatus, kuulua, laitos, maalaisliitto, maaseutu, pienteollisuus, prosentti, vuosikymmen
	Luonnonlaki	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, ryhmä, tietoisuus, tutkimus, yhteiskunta	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, taso, tietoisuus, tutkimus, yhteiskunta	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, taso, tietoisuus, tutkimus, yhteiskunta	ihminen, luonnonlaki, luonto, ohjelma, ongelma, puolue, taso, tietoisuus, tutkimus, yhteiskunta
	Kristilliset	demokraatti, hoito, koulu, kristillinen, lapsi, ohjelma, perhe, päihde, tarve, tuki	demokraatti, hoito, kristillinen, käyttö, lapsi, nuori, parantaa, perhe, saada, tuki	demokraatti, elämä, hoito, kristillinen, lapsi, nuori, perhe, päihde, saada, tuki	demokraatti, euro, hoito, koulu, kristillinen, lapsi, ohjelma, perhe, päihde, tuki
	Piraattipuolue	demokratia, kansalainen, muutos, oikeus, piraatti, puolue, sananvapaus, suomi, tieto, viestintä		demokratia, kansalainen, käyttö, muutos, oikeus, piraatti, puolue, sananvapaus, tieto, vapaus	demokratia, kansalainen, käyttö, oikeus, piraatti, poista, puolue, sananvapaus, tieto, viestintä
	Vihreät	demokratia, eläin, esimerkki, maailma, oikeus, perus, reilu, tukea, unioni, vihreä	energia, kulutus, käyttö, luonto, päästö, suomi, tukea, vero, vihreä, ympäristö	energia, esimerkki, käyttö, luonto, nosta, päästö, suomi, vero, vihreä, ympäristö	energia, kulutus, käyttö, päästö, suomi, tukea, vero, vihreä, ympäristö
	Vihreä unioni	asunto, energia, eurooppa, liitto, luku, luonto, nainen, ongelma, suomi, ympäristö	avoimuus, demokratia, esimerkki, nykyinen, oikeus, reilu, tieto, tulo, unioni, vihreä	demokratia, eläin, esimerkki, globalisaatio, kansalaisuus, nykyinen, perus, reilu, unioni, vihreä	demokratia, eläin, esimerkki, oikeudenmukaisuus, perus, reilu, tukea, tulo, unioni, vihreä
	Sosialismi	kapitalismi, kapitalisti, kehitys, kommunisti, puolue, sosialismi, sosialistinen, taistelu, työväenluokka, yhteiskunta	kansa, kapitalistinen, kehitys, maa, puolue, sosialismi, sosialistinen, suomi, tuotanto, voima	etu, kansa, kapitalistinen, maa, puolue, sosialismi, sosialistinen, suomi, tuotanto, voima	kansa, kapitalistinen, kehitys, maa, puolue, sosialismi, sosialistinen, suomi, tuotanto, voima
	Kokoomus			julkinen, kokoomus, luku, markkinatalous, osake, rahoitus, sosiaalinen, teollisuus, vero, yritys	elinkeinoelämä, julkinen, kansalainen, kokoomus, luku, markkinatalous, osake, valta, vero, yritys

Jatkuu seuraavalla sivulla

Taulukko L1 – jatkuu

	Tulkinta	$k = 25$	$k = 28$	$k = 33$	$k = 34$
	Kokoomus			eettinen, henkinen, kilpailu, kokoomus, moniarvoisuus, valinta, vastuu, yksilö, yksilöllinen, yksilöllisyys	henkinen, kansanvalta, kilpailu, kokoomus, toimenpide, turvallisuus, valinta, yksilö, yksilöllinen, ympäristönsuojelu
	Työväenpuolue	eurooppa, hyvinvointi, kansa, pääoma, suomi, tavoite, työväenpuolue, unioni, vaihtoehto, yhteinen	eurooppa, hyvinvointi, kansa, kunta, pääoma, suomi, tavoite, toiminta, työväenpuolue, yhteinen	hyvinvointi, järjestö, kansa, katsoa, sota, suomi, tavoite, työväenpuolue, yhteinen, yhteiskuntapolitiikka	hyvinvointi, järjestö, kansa, katsoa, nykyinen, sota, suomi, tavoite, työväenpuolue, yhteinen
	Vasemmistoliitto	arvo, kehitys, liitto, markkinat, oikeus, poliittinen, sosiaalinen, vapaus, vasemmisto, yritys	ehdottaa, hyvinvointivaltio, kehitys, liitto, nainen, ohjelma, oikeus, pääoma, suomi, vasemmisto	ehdottaa, hyvinvointivaltio, liitto, nainen, oikeus, palkka, pääoma, sukupuoli, tasainen, vasemmisto	ehdottaa, hyvinvointivaltio, liitto, nainen, ohjelma, oikeus, pitää, pääoma, tasainen, vasemmisto
	RKP	kansanpuolue, kieli, maa, oikeus, pitää, puolue, ruotsalainen, ruotsinkielinen, suomi, väestö	kansanpuolue, kieli, pitää, puolue, ruotsalainen, ruotsi, ruotsinkielinen, suomenruotsalainen, vaatia, väestö	edistä, kansanpuolue, kieli, puolue, ruotsalainen, ruotsi, ruotsinkielinen, suomenruotsalainen, suomi, väestö	edistä, kansanpuolue, kieli, pitää, puolue, ruotsalainen, ruotsi, ruotsinkielinen, suomenruotsalainen, väestö
	Liberaalit	eläke, julkinen, kansalainen, liberaali, oikeus, palvelu, perus, tulo, työ, yksityinen	ehdokas, eläke, julkinen, kansalainen, laki, liberaali, oikeus, perus, tulo, vapaus	eläke, julkinen, kansalainen, laki, liberaali, perus, tulo, vapaus, vero, verotus	eläke, julkinen, kansalainen, liberaali, palvelu, perus, pitää, tehtävä, tulo, vakuutus
	SDP		demokratia, julkinen, kansalainen, sosialidemokraattinen, taloudellinen, tavoite, toteuttaminen, työntekijä, valta, yhteiskunnallinen	muutos, parantaminen, sdp, sosialidemokraattinen, taloudellinen, tasainen, toteuttaminen, työntekijä, työväenliike, uudistaminen	kansanvalta, puolue, sosialidemokraattinen, tavoite, toteuttaminen, työ, työntekijä, työväenliike, valta, yhteiskunnallinen
	SKDL			alista, demokraattinen, kehittäminen, kehitys, liike, skdl, toiminta, työntekijä, vaihtoehto, yhteistyö	demokraattinen, henkinen, laaja, liitto, luoda, maa, poliittinen, skdl, yhteiskunnallinen, yhteistyö
	Perussuomalaiset			kansa, kansalainen, maa, perus, poista, puolue, saada, suomalainen, suomi, turvata	etu, kansa, kansalainen, maa, perus, puolue, saada, suomalainen, suomi, turvata
Politiikka-aiheet					
	Yrittäjyys		kansalainen, kilpailu, koti, nuori, sosiaalinen, teollisuus, tuotanto, valtiovalta, yrittäjä, yritys	edellytys, kansainvälinen, kansalainen, kehittäminen, kehitys, kulttuuri, taloudellinen, toiminta, yhteiskunnallinen, yhteiskunta	kehitys, markkinat, oikeus, poliittinen, sosiaalinen, tuotanto, vapaus, vasemmisto, yhteiskunta, yritys
	Inhimillisyys		huomio, ihminen, ihmisyys, inhimillinen, maailma, mahdollinen, pitää, puolue, vaatia, yhteiskunnallinen	huomio, ihminen, ihmisyys, inhimillinen, maailma, mahdollinen, pitää, puolue, tekijä, vaatia	huomio, ihminen, ihmisyys, inhimillinen, maailma, mahdollinen, oikeuslaitos, puolue, taide, vaatia
	Demokratia	demokraattinen, demokratia, kansa, kehitys, maa, suomi, taloudellinen, tuotanto, yhteiskunnallinen, yhteistyö	demokratia, ihminen, kehitys, markkinat, oikeus, poliittinen, sosiaalinen, vapaus, vasemmisto, yhteiskunta		
	Alueet	ala, alue, asunto, laina, luku, myöntää, perhe, suorittaa, teollisuus, toimenpide	alue, hinta, kaupunki, korko, laina, laitos, maatalous, palkka, prosentti, pula	alue, hinta, kaupunki, laitos, piiri, pohjoinen, seutu, teollisuus, toiminta, vesi	
	Luonto ja ympäristö	alue, energia, ihminen, käyttö, lapsi, liikenne, luonto, saada, suomi, ympäristö	ihminen, jäte, koulu, käyttö, lapsi, metsä, saada, suomi, tavoite, ympäristö	ihminen, kulttuuri, käyttö, lapsi, liikenne, luonto, metsä, suomi, turku, ympäristö	energia, ihminen, jäte, käyttö, lapsi, metsä, saada, suomi, turku, ympäristö
	Alueet			alueellinen, hallinto, huomio, ihminen, kunnallinen, kunta, mahdollisuus, saattaa, tieto, valtio	
	Edut ja turvaaminen	estä, etu, haluta, ihmisyys, kansa, pitää, puolue, saada, työ, vaatia	etu, kansa, kansalainen, maa, poista, puolue, saada, suomalainen, suomi, turvata		

Jatkuu seuraavalla sivulla

Taulukko L1 – jatkuu

	Tulkinta	$k = 25$	$k = 28$	$k = 33$	$k = 34$
	Verotus	eurooppa, kehittää, koulutus, palvelu, suomi, taso, tavoite, tuki, työ, verotus	järjestelmä, koulutus, kunta, palvelu, suomi, taso, tuki, työ, vero, yritys		eurooppa, koulutus, kunta, palvelu, suomi, taso, tuki, työ, verotus, yritys
	Kieli ja koulutus	kieli, kunta, laki, maa, määrätä, oikeus, saada, suomi, valtio, yleinen	koulu, kunta, laki, maa, määrätä, oikeus, saada, suomi, valtio, yleinen	kansa, koulu, kunta, laki, maa, oikeus, saada, suomi, valtio, yleinen	kieli, kunta, laki, maa, määrätä, oikeus, saada, suomi, valtio, yleinen
	Mahdollisuudet	maa, nykyinen, osa, pitää, saada, suomi, tehdä, tärkeä, valtio, voida	maa, mahdollinen, nykyinen, osa, saada, suomi, tehdä, työ, valtio, voida	käyttää, maa, nykyinen, osa, pitää, saada, suomi, tehdä, valtio, voida	järjestelmä, maa, nykyinen, ongelma, osa, saada, suomi, tehdä, valtio, voida
	Edut ja turvaaminen	etu, kansa, köyhälistö, pula, suomi, tahtoa, työ, työttömyys, työväki, yhteinen		etu, kansa, korko, köyhälistö, pula, rintama, tahtoa, työ, työväki, yhteinen	etu, korko, köyhälistö, pula, rintama, tahtoa, työ, työväki, valta, yhteinen
	Hyvinvointipalvelut	edistää, hyvinvointi, kunnallinen, kunta, lapsi, oikeus, palvelu, suomalainen, työntekijä, yksityinen		esimerkki, hyvinvointi, kunta, maksu, malli, osaaminen, palvelu, suomi, tuki, unioni	hyvinvointi, kunnallinen, kunta, kuntalainen, malli, palvelu, tarve, terveys, vanhus, yksityinen
	Arvot	arvo, ihminen, kansalainen, kulttuuri, luonto, mahdollisuus, talous, työ, yhteiskunta, ympäristö	arvo, ihminen, kansalainen, kehitys, mahdollisuus, talous, työ, voida, yhteiskunta, ympäristö	arvo, ihminen, kansalainen, mahdollisuus, oikeus, suomi, talous, työ, yhteiskunta, ympäristö	arvo, ihminen, kansalainen, kehitys, mahdollisuus, oikeus, työ, voida, yhteiskunta, ympäristö
	Maaseudun asema	asema, huomio, kansa, maa, maaseutu, maatalous, sivistys, taloudellinen, työ, valtio	huomio, kansa, kehittää, maa, maaseutu, maatalous, taloudellinen, toiminta, työ, valtio	huomio, kansa, kansalainen, maa, mahdollisuus, pyrkiä, toiminta, työ, valtio, yhteiskunta	asema, huomio, maa, maatalous, saada, tarve, toimenpide, toiminta, työ, valtio
	Kehittäminen ja mahdollisuudet	huomio, kehittäminen, kehittää, koulutus, lisätä, mahdollisuus, taloudellinen, toiminta, turvata, työ	huomio, kehittää, koulutus, kunta, lisätä, maa, mahdollisuus, taloudellinen, toiminta, turvata	järjestelmä, kehittää, koulutus, lisätä, maa, mahdollisuus, tavoite, toiminta, turvata, työ	järjestelmä, kehittäminen, kehittää, koulutus, lisätä, mahdollisuus, taloudellinen, tavoite, toiminta, turvata
	Miehet ja naiset		asunto, eurooppa, mies, nainen, neuvosto, puolue, suomalainen, taloudellinen, vuosikymmen, yhteinen	henkinen, korkea, kulttuuri, lama, mies, nainen, puolue, suomalainen, tuoda, vuosikymmen	asunto, eurooppa, keskeinen, kulttuuri, mies, nainen, taloudellinen, vuosikymmen, yhteinen, ympäristö
	Uskonto		elämä, elää, hengellinen, jumala, koti, kristillinen, kristinusko, tehtävä, tuki, yhteisö		
	Ihmiskunta				alista, eurooppa, ihmiskunta, kapitalismi, kasvaa, markkinat, raha, työttömyys, unioni, vaihtoehto