# Aihemallinnuksesta kehysmallinnukseen

TUUKKA YLÄ-ANTTILA, VEIKKO ERANTI JA ANNA KUKKONEN

## Johdanto

Aihemallinnus (*topic modeling*, ks. esim. Bail 2014; DiMaggio ym. 2013; Evans 2014; Purhonen ja Toikka 2016) on yksi uusista tiedonlouhintamenetelmistä, joita on viime vuosina sovellettu myös yhteiskuntatieteellisessä tutkimuksessa. Näitä big data -menetelmiksikin kutsuttuja lähestymistapoja on leimannut vahva luottamus induktiiviseen logiikkaan, joka perustuu säännönmukaisuuksien etsimiseen suurista aineistoista ilman teoreettisia ennakko-oletuksia, ja usein myös innokas kausaaliväitteiden esittäminen (Babones 2016). Pyrimme lähestymään aihepiiriä hieman toisenlaisesta näkökulmasta: miten sellainen laadullinen tutkimusote, joka lähtökohtaisesti perustuu merkitysten tulkintaan, voi hyötyä aihemallinnuksesta? Miten algoritmisilla menetelmillä voitaisiin löytää kulttuurisia rakenteita teksteistä? Sovellamme tässä kirjoituksessa aihemallinnusta erityisesti kehysanalyysiin, tarkemmin sanoen julkisen keskustelun kehysanalyysiin. Tapausesimerkkinä toimii ilmastonmuutosta käsittelevä keskustelu julkisuudessa – kenttä, jonka tutkimus on pitkään nojannut pääosin laadullisiin menetelmiin, mutta voisi hyötyä tiedonlouhinnasta (Broadbent ym. 2016).

Yhden määritelmän mukaan kehys on ”linkki kahden käsitteen välillä, jolle altistuttuaan yleisö ymmärtää näiden käsitteiden välillä olevan yhteyden” (Nisbet 2009, 17). Toisin sanoen kehystäminen on yksinkertaisimmillaan sitä, kun asian A esitetään olevan merkityksellinen asian B ymmärtämiselle. Tällaisten yhteyksien löytämiseen ovat omiaan sellaiset algoritmiset menetelmät, jotka etsivät käsitteitä, jotka ”tapaavat esiintyä yhdessä tietyissä teksteissä” (DiMaggio ym. 2013, 578). Juuri tätä aihemallit tekevät. Tiettyjen sanojen toistuva ja tavanmukainen käyttäminen yhdessä toistensa kanssa viittaa siihen, että noiden sanojen välillä on merkitysyhteys, ja joukko tällaisia toisiinsa liittyviä sanoja voidaan tulkita jälkeenpäin kehyksestä. Aihemallinnuksella voisi siis kenties automatisoida osan kehysanalyysin prosessista: kehysten tunnistamisen.

Aihemallinnusta ja muita tekstinlouhintamenetelmiä, jotka etsivät säännönmukaisuuksia suurista tekstiaineistoista, voidaan pitää ”etäluentana” (Moretti 2013), vastakohtana laadullisten tekstianalyysimenetelmien ”lähiluennalle”. Aihemallinnus redusoi kielen monimutkaiset vivahteet äärimmäiseen yksinkertaistukseen: tietyt sanat esiintyvät usein yhdessä, ja näiden sanojen joukot sekä sanajoukkojen esiintymistiheys ja sen vaihtelu ovat jälkiä kulttuurisista merkityksistä. Aihemallinnuksella on mahdollista ainakin pintapuolisesti analysoida suurempia tekstiaineistoja kuin lähilukemalla ja löytää sellaisia merkitysrakenteita, joiden olemassaoloa tutkija ei välttämättä olisi lukemalla huomannut. Silti merkitysten *ymmärtämisen* kannalta lähiluenta on yhä ensiarvoisen tärkeää. Aihemallinnuksen hyödyntäminen yhteiskuntatieteellisessä tekstien tutkimuksessa on siis väistämättä enemmän tai vähemmän monimenetelmätutkimusta.

Esimerkkiaineistomme koostuu ilmastonmuutosaiheisista kirjoituksista kahdessa englanninkielisessä lehdessä, *New York Timesissa* (USA) ja *The Hindussa* (Intia). Tutkimme, miten eri toimijat kehystävät ilmastonmuutosta julkisuudessa näissä kahdessa eri maassa.

Ilmastonmuutoksesta käydyssä julkisessa keskustelussa esiintyy tavattoman paljon erilaisia käsityksiä itse ongelman luonteesta ja siihen sopivista ratkaisuista. Suuri joukko erilaisia toimijoita, kuten valtioita, järjestöjä ja asiantuntijaorganisaatioita, kilpailevat siitä, miten ilmastonmuutosta kehystetään mediassa (Anderson 2009, Boykoff 2011; Hajer 1995; Nisbet 2009; Schäfer ja Schlichting 2014; Trumbo ja Shanahan 2000). Tämä johtuu siitä, että ilmastonmuutosta tulkitaan erilaisten arvojen ja uskomusten pohjalta (Boykoff ja Crow 2014; Hoffman 2015; Hulme 2009).

Esimerkkianalyysimme pohjalta esitämme, että kehysanalyysin kehykset voidaan operationalisoida aihemallinnuksen aiheina (Bail 2014; DiMaggio ym.

2013), jos

1. kehykset määritellään kahden käsitteen välisiksi yhteyksiksi (Entman 1993; Nisbet 2009)
2. aineistona käytetään tekstejä, jotka käsittelevät jotain tiettyä aihetta
3. malli validoidaan kehysanalyyttisesti.

Esitämme näiden ehtojen toteuttamiseksi käytännön ohjenuoria. Esimerkkianalyysimme perusteella näyttää, että taloudellinen kehystäminen on tavallisin tapa ymmärtää ilmastonmuutosta *New York Timesissa*, kun taas *The Hindussa* korostetaan vastuunjakoa ja ympäristöriskejä. Jotkut kehykset jakavat keskusteluun osallistuvia ryhmiä enemmän kuin toiset: esimerkiksi kansalaistoiminnasta puhuvat lähinnä järjestöt, ilmastotieteestä lähinnä asiantuntijat. Talouskasvun ja ympäristönsuojelun yhdistävään vihreän kasvun kehykseen eri ryhmät sen sijaan osallistuivat tasaisemmin.

## Aihemallinnus

Käytämme aihemallinnukseen MALLET-ohjelmiston (Machine Learning for Language Toolkit, http://mallet.cs.umass.edu/) sisältämää Latent Dirichlet Allocation -mallia (LDA; Blei ym. 2003), yleisintä aihemallinnuksessa käytettyä mallia. Se on ns. ei-ohjattu (*unsupervised*) koneoppimismalli: sen käyttäjä, ihminen, ei anna mallille itse aineiston ja toivotun aiheiden lukumäärän lisäksi mitään vaatimuksia siitä, miten malli aineistoa luokittelee. Luokitukset (”aiheet”) ovat siis lähes täysin induktiivisia eli ”teoriavapaita” ja perustuvat vain siihen oletukseen, että sanojen esiintyminen yhdessä viittaa niiden väliseen käsitteelliseen yhteyteen. Uskomme kuitenkin, että tulkinnalle jää aihemallinnuksessa jotakuinkin yhtä paljon tilaa kuin perinteisemmässä laadullisessa luennassa; se vain tapahtuu luokitteluvaiheen jälkeen. ”Induktiivisen” tai ”deduktiivisen” logiikan sijaan tällainen koneluokittelun ja ihmistulkinnan yhdistelmä onkin lähempänä abduktiivista prosessia (Glaser ja Strauss 1967; Timmermans ja Tavory 2012).

LDA-aihemallinnus ”olettaa, että tekstikokoelmassa esiintyy joukko aiheita”, että ”sanat, jotka ovat merkittäviä kullekin aiheelle, tapaavat esiintyä yhdessä useammin kuin sattumanvaraisesti”, ja että ”kukin teksti muodostuu näistä aiheista eri suhteissa.” (DiMaggio ym. 2013, 577–578.) LDA on siis tilastollinen malli, joka mallintaa kunkin aiheen eli sanajoukon esiintymistodennäköisyyttä kussakin tekstissä ja kunkin sanan esiintymistodennäköisyyttä kussakin aiheessa eli sanajoukossa. ”Aiheet” ovat sanafrekvenssien todennäköisyysjakaumia (Blei ym. 2003). Yksinkertaisimmillaan tutkija syöttää ohjelmaan aineiston, joka koostuu teksteistä, ja pyytää ohjelmaa palauttamaan tietyn määrän aiheita, vaikkapa kymmenen. Ohjelma tulostaa 1) kymmenen ”aihetta” eli listaa sanoja, jotka usein esiintyvät yhdessä, 2) kunkin tekstin sisältämien aiheiden jakauman, joka siis näyttää, kuinka suuri osuus ko. tekstistä koostuu kustakin ”aiheesta” eli sanaryhmästä, ja 3) kunkin aiheen sisältämien tekstien jakauman, joka näyttää, kuinka suuri osuus ko. ”aiheesta” koostuu kustakin tekstistä. Nämä jakaumat tavallisesti tulkitaan kuvauksiksi siitä, 1) mistä ”aiheista” aineisto ”kertoo”, 2) mistä ”aiheista” kukin teksti ”kertoo”, ja 3) missä teksteissä kutakin ”aihetta” käsitellään ja minkä verran.

## Kehysmallinnus

Sana ”aihe” ei siis oikeastaan viittaa mihinkään itse mallin ominaisuuteen, koska malli ”ymmärtää” vain sanojen (tai oikeastaan minkä tahansa tunnistettavien yksiköiden) yhteisesiintymiä. Näiden kutsuminen ”aiheiksi” on jo operationalisointi, tulkinta siitä mihin mallia voidaan käyttää. Alun perin esitellessään LDA-mallia David Blei ja kollegat kirjoittivat käyttävänsä artikkelissaan ”tekstikokoelmiin liittyvää kieltä, siis ’sanoja’, ’tekstejä’ ja ’kokoelmia’, ymmärryksen ohjaamiseksi” (Blei ym. 2003, 995). Sana ”aihe” on ”ohjannut ymmärrystä” kenties liikaakin (esim. LDA:n soveltamisesta laivojen reittien luokitteluun ks. Schmidt 2012). Sellaisessa tutkimusasetelmassa, jossa aineistona käytetään tekstejä, joiden jo tiedämme käsittelevän tiettyjä temaattisia aiheita (kuten ilmastonmuutosta), LDA:n voidaan tulkita mallintavan pikemminkin t*apoja puhua tietystä aiheesta*, siis kehyksiä. Se operationalisoi abstraktimman tutkimuskysymyksen ”miten tästä aiheesta puhutaan” konkreettisemmaksi analyyttiseksi kysymykseksi ”millä sanoilla tästä aiheesta puhutaan”. Aihemallinnuksen yhteiskunnalliset sovellukset usein painottavat tulosten validointia (DiMaggio ym. 2013; Evans 2014; Grimmer ja Stewart 2013), siis sen varmistamista, että mallin sanajoukot kuvaavat sitä mitä luulemme niiden kuvaavan. Jos olemme kiinnostuneita kehyksistä pikemminkin kuin aiheista, tämän validoinnin tulee tapahtua kehysanalyysin puitteissa.

Ensinnäkin, kehysmallin sisäisen validiteetin kan-

nalta sanajoukkojen täytyy muodostaa koherentteja ”tulkinnallisia skeemoja”, jotka ”auttavat käyttäjiään paikantamaan, havaitsemaan, tunnistamaan ja nimeämään” kokemuksia ja tapahtumia (Goffman 1974, 21). Erving Goffman loi kehyksen käsitteen kuvaamaan kasvokkaista vuorovaikutusta, mutta sen jälkeen kehystämistä käsittelevä kirjallisuus on kehittynyt moniin eri suuntiin. Tästä valtavasta kirjallisuudesta valitsemme kuitenkin Robert Entmanin (1993) yksinkertaistetun määritelmän, koska se sopii tarpeisiimme: “Kehystäminen on sitä, kun koetusta todellisuudesta valitaan joitain ominaisuuksia ja tehdään niistä erityisen keskeisiä tavalla, joka edistää tiettyä määritelmää ongelmasta, tulkintaa syy-seuraussuhteista, moraalista arvostelmaa ja/tai ratkaisuehdotusta.” (Entman 1993, 52.) Kehystämisessä siis esitetään jonkin asian olevan erityisen tärkeä jonkin toisen asian tulkitsemiselle.

Kehykset ”määrittelevät ongelmia”, ”diagnosoivat syitä”, ”ottavat moraalisia kantoja” ja ”ehdottavat ratkaisuja”, ja ne tapahtuvat ”viestintäprosessissa” (Entman 1993, 52). On kuitenkin tärkeää huomata, että ”kehyksiä ei pidä sekoittaa tiettyihin poliittisiin kantoihin; jokin tietty kehys voi sisältää niin neutraaleja argumentteja kuin argumentteja puolesta tai vastaan” (Nisbet 2009, 18). Kehys siis määrittelee, millä seikoilla on väliä käsillä olevalle asialle.

Toiseksi, kehysmallinnuksen ulkoinen validiteetti tarkoittaa sitä, että löydettyjen kehysten pitäisi jotakuinkin vastata olemassa olevaa tietoa kyseisestä tutkimuskohteesta, jotta malliin voidaan luottaa. Samalla kuitenkin täytyy jättää tilaa uusille havainnoille, tai aihemallinnuksen hyöty kyseenalaistuu. Mallihan ei ole kovin uskottava, jos sitä käyttämällä saadaan tuloksia, jotka asettuvat kaikkea aiempaa, muilla menetelmillä tehtyä tutkimusta vastaan. Sen sijaan yhteensopivuus mallista tehtyjen löydösten ja aiemman tutkimuksen välillä vahvistaa samasta mallista tehtyjen uusien, kenties yllättävienkin löydösten uskottavuutta. Siis jos esimerkkitapauksessamme löydökset vastaavat ilmastonmuutoskeskustelusta aiemmin löydettyjä kehyksiä (kuten taloudellinen kilpailukyky, moraali sekä tieteellinen epävarmuus; ks. Nisbet 2009, 18), malli saa vahvistusta. Mallimme löytää kuin löytääkin esimerkkejä näistä kehyksistä, mutta spesifimmässä muodossa.

## Esimerkkiaineisto

Mediajulkisuus on tärkeä areena ilmastonmuutoksesta käytävälle poliittiselle keskustelulle, jossa eri osanottajat kehystävät aihetta eri tavoin (Boykoff 2011; Nisbet 2009). Aiempaa tutkimusta yksittäisten maiden mediakeskusteluista on paljon, ja vertailevissa tutkimuksissa on todettu, että keskustelujen sisältö vaihtelee kulttuurisesta kontekstista toiseen (Broadbent ym. 2016; Kukkonen ym. 2018; Painter ja Ashe 2012; Schmidt ja Schäfer 2015; Ylä-Anttila ja Kukkonen 2014). Aihetta on tutkittu erityisesti Yhdysvalloissa, jossa ilmastonmuutosuutisoinnin on havaittu korostavan tieteellisen tutkimuksen epävarmuutta ja ilmastonmuutoksen torjunnan kustannuksia – ja konservatiivisen liikkeen rooli näissä keskusteluissa on vahva (Dunlap ja McCright 2015; Farrell 2015, 2016; Hoffman 2011; McCright ja Dunlap 2003; Oreskes ja Conway 2010). Vastapainoksi olemme valinneet tutkimusaineistoa Intiasta, jossa ilmastonmuutosta koskeva kirjoittelu on aiempien havaintojen mukaan keskittynyt ympäristöriskeihin ja ilmastopolitiikan kansainväliseen ulottuvuuteen, kuten vastuunjakoon pohjoisen rikkaiden ja etelän köyhien maiden välillä (Billet 2010).

Aiempi tutkimus on siis pääasiassa keskittynyt analysoimaan kehystämistä yhden maan median sisällä, laajemmin kehystämisen eroja maiden välillä tai keskittynyt yhteen toimijatyyppiin (Broadbent ym. 2016; Farrell 2015, 2016; Nisbet 2009). Eri puhujaryhmien, kuten valtioiden, järjestöjen ja asiantuntijoiden, eroja kehystämisessä on tutkittu vähemmän.

Käytämme aineistoa, johon on kerätty kaikki *New York Timesissa* ja *The Hindussa* tiettynä aikana julkaistut artikkelit, joissa mainitaan ilmastonmuutos (”*climate change*” tai ”*global warming*”). Molemmat edustavat maansa liberaalia, laajalevikkistä valtavirtalehdistöä. Aineisto kattaa kunkin lehden osalta kolmen viikon ajanjakson ennen kolmea kansainvälistä ilmastokokousta (Kioto 1997, Kööpenhamina 2009 ja Durban 2011) sekä kolme viikkoa kunkin kokouksen jälkeen. Aiempaa tutkimushanketta (Ylä-Anttila ym. 2018) varten aineistoon oli Atlas.TI-ohjelmalla käsin merkitty poliittiset vaateet, ja ne oli luokiteltu puhujaryhmittäin (asiantuntija, valtio tai järjestö). Poliittinen vaade on ”toiminnan yksikkö julkisessa sfäärissä. Vaade voi olla kommentti haastattelussa tai puheessa, mielenosoitus tai muu teko, jonka tarkoitus on vaikuttaa julkiseen keskusteluun.

Yksi lehtiartikkeli voi siis sisältää useita vaateita useiden toimijoiden esittäminä" (Ylä-Anttila ja Kukkonen 2014, 398–399). Valtio-kategoria sisältää valtioiden edustajien lisäksi valtioiden väliset järjestöt kuten Yhdistyneet Kansakunnat. Puhuja koodattiin asiantuntijaksi, jos tämä esitti itsensä asiantuntijaorganisaation kuten yliopiston edustajana. Järjestö-kategoria kattaa kansalaisjärjestöt, esim. Greenpeace ja muut ympäristöjärjestöt.

**Taulukko 1. Esimerkkiaineisto.**

| | Artikkeleita | Vaateita |
|---|---|---|
| **New York Times** | 94 | 353 |
| **The Hindu** | 583 | 383 |
| **Yhteensä (sanoja)** | 677 (416 822) | 736 (103 589) |

Esitimme aiemmin, että LDA-mallin tulkitsemiseksi kehysmallina käytetyn aineiston täytyy mahdollisimman tarkasti käsitellä tiettyä aihetta. Siispä syötimme malliin vain aiemmin tunnistetut poliittiset vaateet koko lehtiartikkelien sijaan tarkentaaksemme aineistoa entisestään, niin että se sisältää vain poliittisia vaateita, ei niiden ympärillä lehtiartikkeleissa esiintyvää kuvailevaa tekstiä. Muitakin menetelmiä voitaisiin käyttää aineiston valikoimiseksi: esim. Karen Levy ja Michael Franklin (2013) käyttivät aihemallinnuksen aineistona kommentteja kuorma-autoalan sääntelyä koskevaan debattiin. Tärkeintä on temaattinen rajaus, jos kiinnostuksemme kohteena ei ole temaattinen vaihtelu vaan se, miten tiettyä teemaa puheenvuoroissa käsitellään.

Vaateet siirrettiin Atlas.TI:stä tekstitiedostoihin, jotka nimettiin puhujaryhmän, maan ja tunnistenumeron mukaan. Ne *tokenisoitiin* ja *stemmattiin* käyttäen yksinkertaisia Python-ohjelmointikielellä kirjoittamiamme ohjelmia ja Snowball-kirjastoa (http://snowballstem.org/): tekstitiedostoista siis poistettiin välimerkit ja järjesteltiin ne niin, että kukin sana on omalla rivillään, ja sanoista poistettiin taivutuspäätteet, jotta ohjelma tunnistaisi saman sanan eri muodot samaksi sanaksi (esim. "*changed*", "*changing*" ja muut "*change*"-sanan eri taivutusmuodot muutettiin muotoon "*chang*"). Lopulta käytimme MALLETin englanninkielistä *stopword*-listaa, joka poistaa tiedostoista mm. konjunktiot ja muut sellaiset sanat, joilla ei ole tulkittavissa olevaa merkityssisältöä itsessään.

## Kehyksiä mallista

Mallin validointi ja tulkinta mainitaan usein keskeisenä ongelmakohtana käytettäessä aihemallinnusta yhteiskunnallisten ilmiöiden tutkimukseen (DiMaggio ym. 2013; Evans 2014; Grimmer ja Stewart 2013). Tukeaksemme väitettämme, jonka mukaan LDA-mallin löydökset voivat olla tulkittavissa kehyksinä, esitämme kolmivaiheisen prosessin kehysten validointiin. Ensimmäinen vaihe tarkastelee mallia pintapuolisesti kokonaisuutena, toisessa luemme kunkin sanaryhmän yleisimmät sanat tarkistaaksemme mallin sisäisen validiteetin (sanaryhmän sisäinen koherenssi), ja kolmannessa vaiheessa luemme itse aineistosta kunkin aiheen tärkeimmät tekstit tarkistaaksemme ulkoisen validiteetin (olemmeko löytäneet samoja kehyksiä kuin muut, eri menetelmiä käyttäneet tutkijat).

Koska LDA:n tapauksessa ohjelmalle ei anneta muuta syötettä kuin itse aineisto ja haluttu sanaryhmien lukumäärä, tämä luku on erityisen tärkeä ja vaikuttaa paljonkin siihen, miten hyvin malli sopii aineistoon (Evans 2014). Kehysten lukumäärän valinta onkin ensimmäinen askel mallin validoinnissa. Liian monen kehyksen malli löytää vähäpätöisiä, pikkutarkkoja kehyksiä, kun taas liian pieni kehysten määrä johtaa usean kehyksen sekoittumiseen yhteen. Tätä esimerkkianalyysia varten kokeilimme lukumääriä kymmenestä sataan, kymmenen kehyksen välein ja tarkastelimme kunkin mallin kohdalla kunkin sanaryhmän kymmenen yleisimmän sanan listaa etsiäksemme sanaryhmiä, jotka näyttivät ilmastonmuutoskehyksiltä. Päädyimme 30 kehyksen malliin. Tässä vaiheessa on parempi valita liikaa kuin liian vähän kehyksiä, sillä aiheeseen liittymättömät kehykset voidaan myöhemmin hylätä, kun taas useampaa kehystä sekoittavaa sanaryhmää ei voi myöhemmin jakaa moneen osaan. Jo tässä vaiheessa on huomattava, että mallintamiseen sisältyy runsain mitoin subjektiivista tulkintaa, jossa on tärkeää tuntea tutkimansa aihe ja aineisto. Mallin sovittamisesta voidaan toki laskea myös tunnuslukuja (ks. esim. Chang ym. 2009), mutta niiden hyöty on rajallinen, sillä mitä enemmän kehyksiä malliin sisällyttää, sitä tarkemmin se toki periaatteessa kuvaa aineistoa, mutta sitä vähemmän mallintamisesta on hyötyä verrattuna lähiluentaan, kun tulkintatyön määrä kasvaa (Ylä-Anttila 2018). Tarkoituksena on kuitenkin kuvata aineisto yksinkertaistetussa muodossa, ei liian tarkasti.

Toisessa vaiheessa luimme valitsemamme kolmenkymmenen kehyksen mallin kolmestakymmenestä sanalistasta kymmenen ensimmäistä sanaa kustakin tarkistaaksemme niiden sisäisen validiteetin – linkittyikö niissä ilmastonmuutos koherenttiin joukkoon muita käsitteitä (Nisbet 2009)? Kymmenen ”ensimmäisen” sanan logiikka piilee siinä, että LDA:ssa kunkin sanan ei tarvitse kuulua vain yhteen sanaryhmään (tässä kehykseen) vaan useampaan, mutta kuhunkin eri todennäköisyyksillä. MALLET tulostaa kunkin aiheen kohdalla listan sanoja frekvensseineen. Yleisimmät sanat saavat sanaryhmissä suurimman painoarvon. Tässäkin karsintavaiheessa aihetuntemus on tärkeää. Hylkäsimme 30 kehyksen mallista 13 ja pidimme 17 kehystä, joille kullekin annoimme alustavan, sanajoukkoa kuvaavan nimen. Esimerkki hylätystä kehyksestä, joka ei meistä vaikuttanut sisäisesti koherentilta, oli ”*concern, clear, don't, give, document, tax, accept, base, thing, main*”; emme osanneet tulkita, mistä tässä kehyksessä olisi kyse. Esimerkkinä koherentista kehyksestä toimii ”*warm, global, scientist, research, univers(e/al/ity), atmospher(e), caus(e), stud(y/ies), effect, release*”, jonka tulkitsimme edustavan kehystä nimeltä ”ilmastotiede”.

**Taulukko 2. Ilmastonmuutoskehykset ja kunkin kymmenen tärkeintä sanaa New York Timesissa ja The Hindussa.**

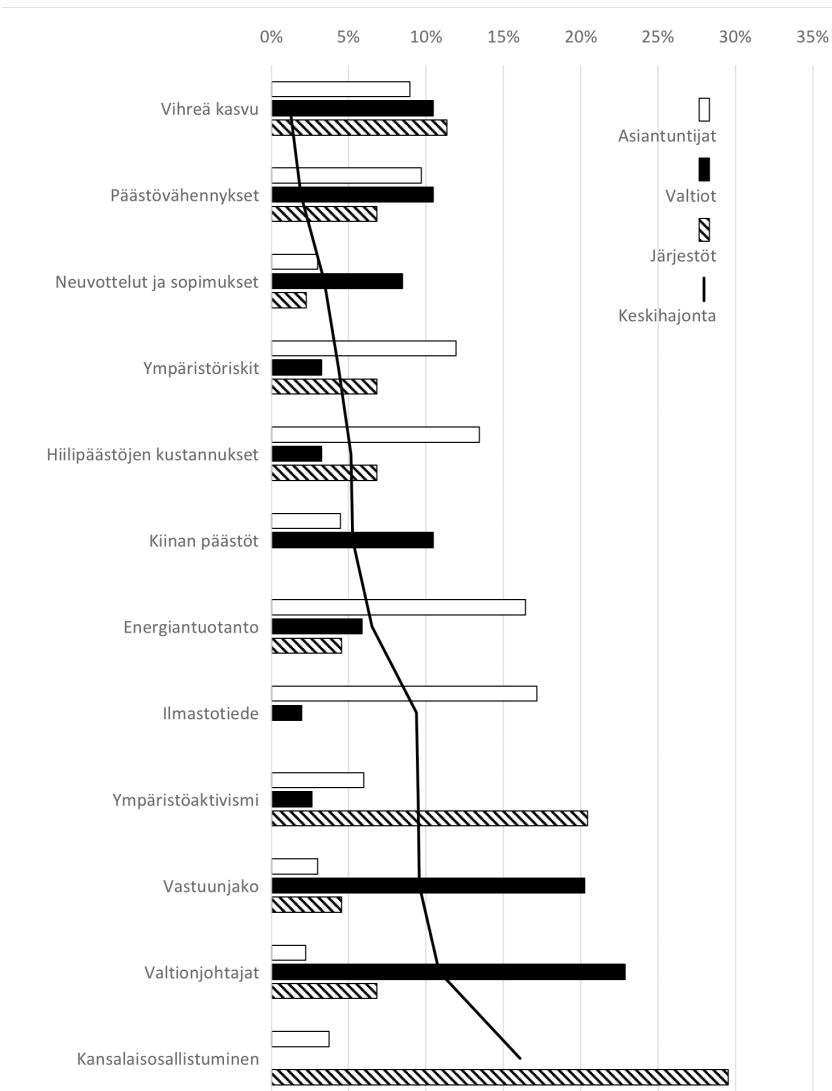
| **Vihreä kasvu** | **Päästövähennykset** | **Neuvottelut ja sopimukset** | **Ympäristöriskit** |
|---|---|---|---|
| energi | emiss | nation | indian |
| fund | cut | unit | sea |
| billion | greenhous | state | water |
| state | industri | treati | increas |
| public | gas | commit | forest |
| clean | gase | economi | today |
| invest | call | major | ocean |
| renew | system | respons | region |
| creat | adopt | american | risk |
| govern | declar | recent | rate |
| **Hiilipäästöjen kustannukset** | **Kiinan päästöt** | **Energiantuotanto** | **Ilmastotiede** |
| carbon | china | econom | warm |
| emiss | target | technolog | global |
| reduc | reduct | cost | scientist |
| dioxid | chines | compani | research |
| pollut | growth | energi | univers |
| trade | intens | fuel | atmospher |
| cap | current | power | caus |
| coal | reduc | money | studi |
| power | plan | price | effect |
| product | oblig | mani | releas |
| **Ympäristöaktivismi** | **Vastuunjako** | **Valtionjohtajat** | **Kansalaisosallistuminen** |
| peopl | countri | meet | part |
| govern | develop | minist | make |
| environ | talk | confer | citi |
| environment | commit | mr | organis |
| protect | provid | day | green |
| campaign | financ | prime | member |
| tree | demand | singh | differ |
| speak | adapt | thursday | number |
| everi | ensur | announc | initi |
| human | warsaw | attend | greenpeac |

Kolmannessa vaiheessa luimme kymmenen ensimmäistä tekstiä (vaadetta) niistä 17 kehyksestä, jotka läpäisivät edellisen vaiheen, ja vertasimme näitä vaateita siihen nimeen, jonka olimme kehykselle edellisessä vaiheessa antaneet. Jos antamamme nimi kuvasi ainakin kahdeksaa kymmenestä tekstistä, pidimme kehyksen mukana analyysissa. Usein nimi kuvasi tekstejä pienin muutoksin – itse tekstien lukeminen siis auttoi tulkitsemaan kehystä paremmin kuin yleisimpien sanojen. Tässä vaiheessa hylkäsimme kuitenkin vielä viisi kehystä, joiden ensimmäiset sanat olivat kyllä edellisessä vaiheessa näyttäneet muodostavan koherentin kokonaisuuden, mutta harvempi kuin kahdeksan kymmenestä tärkeimmästä tekstistä tosiasiassa vastasi kehykselle antamaamme nimeä. Lopullisessa mallissa (taulukko 2) on siis kaksitoista nimettyä ja eri tavoin tarkistettua kehystä.

## Mallista tulkintaan

Seuraavaksi esitämme lyhyesti löytämämme 12 kehystä. Kuten sanottua, tulkitsemme kehyksiä eri käsitteiden välisinä yhteyksinä: ne ovat esitettyjä näkemyksiä siitä, millä muilla käsitteillä on merkitystä suhteessa ilmastonmuutokseen. Siispä esimerkiksi päästövähennykset-kehys sisältää sekä vaateita, joiden mukaan päästövähennykset ovat tarpeellisia, että vaateita joiden mukaan eivät – mutta kaikki tämän kehyksen vaateet esittävät päästövähennykset keskeisenä ilmastonmuutokseen liittyvänä seikkana. Lainauksissa olemme esittäneet kunkin kehyksen tunnussanat (tärkeimmät 10 sanaa) lihavoituina.

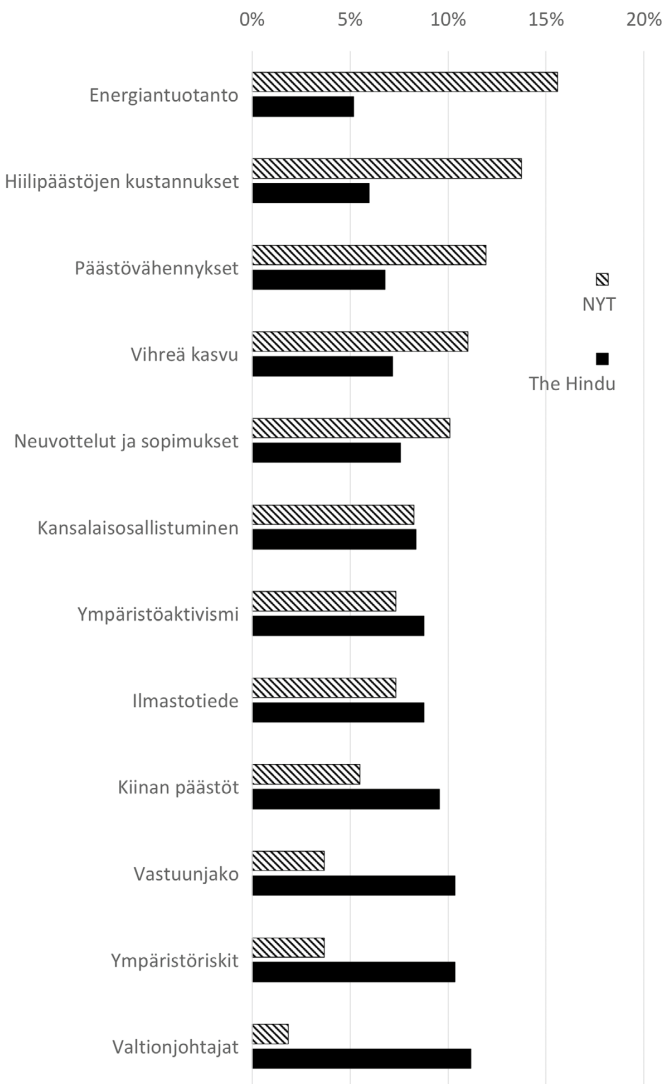
**Kuvio 1. Kehykset puhujaryhmittäin (kunkin puhujaryhmän summa 100%).**




MALLET tulostaa kullekin näistä kehyksistä myös listan, joka kertoo, mistä teksteistä kukin kehys muodostuu. Vertailemalla puhujaryhmien ja maiden osuuksia kunkin kehyksen osalta tässä listassa näemme, mitkä puhujaryhmät useimmin esittivät tähän kehykseen kuuluvia vaateita ja minkä maan lehdistössä vaateet esiintyivät (kuviot 1 ja 2).

Vihreä kasvu -kehys muodostuu sanoista kuten rahoittaa, miljardi, investoida, puhdas ja uusiutua/uusiutuva. Nämä vaateet käsittelevät talouskasvun ja ympäristönsuojelun yhteensovittamista. Tässä kehyksessä jakauma eri

**Kuvio 2. Kehykset maittain (kunkin maan summa 100%).**




puhujaryhmien välillä oli kaikkein tasaisin. Toisin sanoen vihreän kasvun kautta ilmastonmuutosta kehystivät niin valtiot, järjestöt kuin asiantuntijatkin.

> *Using our national development finance institution and export credit agency, we have channelled hundreds of millions of dollars to strengthen India's ability to build technical capacity, reduce financial risk, and lower the cost of capital for low-carbon* ***investments****.* (india-gov247)

Päästövähennykset-kehys koostuu sanoista kuten päästö, leikata, ja kasvihuone(kaasu/ilmiö). Myös tämä kehys jakautui eri puhujaryhmien välille tasaisesti. Neuvottelut-kehys taas kuvaa ilmastoneuvotteluja mm. sanoin kansakunta, valtio, sopimus jne. ja oli luonnollisestikin valtiotoimijoiden dominoima. Ympäristöriskit-kehys koostuu esim. sanoista meri, vesi, lisääntyä, metsä ja riski ja oli huomattavasti yleisempi *The Hindussa* kuin *New York Timesissa*. Hiilidioksidipäästöjen kustannuksia käsittelevä kehys puolestaan käsittää sanat hiili, päästö, kauppa jne. ja se oli erityisen suosittu asiantuntijoiden puheenvuoroissa *New York Timesissa*.

> *But it would be even better, Dr. McKitrick says, to use the temperature readings as the basis for a* ***carbon*** *tax instead of a* ***cap****-and-****trade*** *system.* (usa-expert67)

Kiinan päästöt -kehyksessä käytetään mm. sanoja Kiina, tavoite, kasvu ja vähentää. Jotkut puhujat syyttivät Kiinaa ja vaativat sen kantavan vastuunsa ilmastonmuutoksesta, kun taas toiset ylistivät Kiinaa investoinneista uusiutuvaan energiaan. Energiantuotanto-kehyksessä puhutaan sanoilla talous, kustannus, energia, polttoaine ja raha. Varsinkin asiantuntijat osallistuivat tähän kehykseen, jossa käsiteltiin vaihtoehtoisia energiantuotannon tapoja, erityisesti suhteessa niiden kustannuksiin – mikä selittänee tämän kehyksen runsasta esiintymistä *New York Timesissa*, ottaen huomioon talousargumenttien vahvan aseman amerikkalaisessa poliittisessa kulttuurissa (Lamont ja Thévenot 2000; Rabe 2010).

> *The report said the country was brimming with "negative cost opportunities" – potential changes in the lighting, heating and cooling of buildings, for example, that would reduce carbon dioxide emissions from the burning of fossil* ***fuels*** *even as they save* ***money****.* (usa-expert114)

Ilmastotiede-kehystä määrittävät sanat kuten tutkija ja tutkimus, ja se koostuukin pääosin ilmastotutkimuksesta kertovista uutisjutuista. Ymmärrettävästi se oli vahvasti asiantuntijoiden hallussa. Ympäristöaktivismi-kehys, koostuen sanoista kuten ihminen, ympäristö ja suojelu, oli vastaavasti kansalaisjärjes-

töjen suosiossa. Valtiot taas käyttivät usein vastuunjako-kehystä, jossa kehystetään ilmastonmuutosta sanoilla kehitys/kehittyvä, maa ja rahoitus, sekä valtionjohtajat-kehystä, jossa puhutaan tapaamisista, ministereistä ja kokouksista. Kansalaisosallistuminen-kehys käsittelee kansalaisten roolia ilmastoneuvotteluissa sanoilla osallistua, järjestö, jäsen jne. Tämä kehys oli puhujiltaan yksipuolisin: sitä käyttivät lähes yksinomaan kansalaisjärjestösektorin toimijat.

> *To create public awareness about how climate change is affecting our planet, bicycle enthusiasts took* ***part*** *in a cycle rally on the Capital's much talked about Bus Rapid Transit (BRT) corridor over the weekend.* (india-ngo17)

## Keskustelua

Tämän kirjoituksen tavoitteena oli tarkastella, voidaanko LDA-mallia (yleensä ”aihemallinnusta”) käyttää kehysanalyysin apuvälineenä. Esimerkkiaineistomme käsittää ilmastonmuutoksesta käytyä mediakeskustelua Yhdysvalloissa ja Intiassa vuosien 1997–2011 aikana. Löysimme LDA-mallin avulla 12 sanaryhmää, jotka tulkitsimme kehyksiksi käyttäen esittelemäämme tulkinnan ja validoinnin prosessia. Esimerkkiaineistomme oli laskennalliseksi tutkimukseksi pienehkö, mutta tulkintatyöhön käytetty aika olisi pysynyt pitkälti samana, vaikka aineiston kokoa olisi suurennettu moninkertaiseksi. Tutkimalla kehysten jakaumia puhujaryhmittäin ja maittain havaitsimme eroja intialaisen ja amerikkalaisen sanomalehden tavoissa kehystää ilmastonmuutosta, kuten talousargumenttien vahvan aseman Yhdysvalloissa ja ympäristöriskien painotuksen Intiassa. Jotkut kehykset olivat vahvasti tiettyjen puhujaryhmien hallussa, kuten ilmastotiede asiantuntijoiden. Löysimme kuitenkin myös kehyksiä, jotka jakautuivat eri puhujaryhmien kesken tasaisemmin ja kenties tarjoavat mahdollisuuksia poliittisille kompromisseille, esimerkkinä vihreän kasvun kehys.

Tulkittaessa koneoppimismallien, kuten LDA:n, löydöksiä kulttuurisina konstruktioina, kuten kehyksinä, on otettava huomioon muutama seikka. Ensinnäkin käytimme Robert Entmanin (1993) ja Matthew Nisbetin (2009) yksinkertaistettua kehyksen määritelmää, joka on huomattavasti suurpiirteisempi kuin esim. sosiaalisten liikkeiden tutkimuksessa usein käytetty, strategista toimintaa painottava kehystämisen koulukunta (esim. Benford ja Snow 2000) tai Goffmanin (1974) mikrointeraktioiden tutkimus. Näiden ja muiden teorioiden hienovaraiset vivahteet olisivat vaikeita tai mahdottomia operationalisoida näin yksinkertaisilla työkaluilla. Entmanin ja Nisbetin melko suoraviivaiseen kehyskäsitykseen LDA kuitenkin sopii melko hyvin, kun tarkastelun kohteena ovat sanojen yhteydet toisiinsa.

Toiseksi, mikäli olemme kiinnostuneita puheen tavoista, aineiston tulee olla hyvin tarkasti valikoitu niin, että se sisältää tekstiä vain tietystä aiheesta. Tällöin aihemallinnuksen logiikka muuttuu aihemallinnuksesta kehysmallinnukseksi. Erilaisia mahdollisia tapoja saavuttaa tällainen tiettyä aihetta koskeva aineisto ovat ainakin perinteiset asiasanahaut, mutta miksei myös aihemallinnus itse: sekalaisemmalle aineistolle voisi käyttää ensin yhtä LDA-mallia erottamaan siitä tiettyä aihetta koskeva sisältö, ja sitten ajaa näin erotellulle aineistolle toinen LDA-malli, jota käytettäisiin ilmaisun tapojen erittelemiseen.

Lopuksi, ehdotimme tässä katsauksessa kolmivaiheista tulkinnallisen validoinnin prosessia, joka ensin tarkastelee mallia kokonaisuutena, sitten kehysten sisäistä koherenssia sanalistojen perusteella ja lopuksi kehysten ulkoista validiteettia vertaillen itse tekstejä aiempaan tutkimukseen. Löytämämme kehykset pitkälti vastasivat aiemmin tunnistettuja ilmastonmuutoskeskustelun kehyksiä: löysimme useamman hieman vivahteiltaan poikkeavan kehyksen taloudellisista kustannuksista (hiilipäästöjen kustannukset ja energiantuotanto), moraalisen vastuunjako-kehyksen, sekä ilmastotiede-kehyksen, joka sisälsi tieteellistä epävarmuutta käsitteleviä puheenvuoroja (Nisbet 2009). Nämä erilaiset kehystämisen tavat vaikuttavat siihen, millaisena poliittisena ongelmana ilmastonmuutosta pidetään ja miten se tulisi ratkaista. Eri menetelmin löytyy siis yksityiskohdiltaan hieman erilainen jaottelu, mutta pääpiirteittäin tulokset ovat linjassa aiemman tutkimuksen kanssa. Näin aihemallinnus voi tarjota yhden uuden työkalun kehysten tunnistamiseen.

Emme siis esitä LDA-mallin tai minkään muunkaan algoritmin korvaavan tulkitsevaa kehysanalyysia, vaan toimivan yhtenä apuvälineenä sille. Se voi auttaa tunnistamaan kehyksiä mahdollistaen suurempien aineistojen käytön ja aiemmin huomiotta jääneiden kehysten löytämisen. Tulkinta, kehysanalyysin tärkein osuus, jää kuitenkin edelleen tutkijan itsensä harteille. Automaatio ei ainakaan vielä vie työpaikkojamme, vaan auttaa meitä tekemään työmme tehokkaammin, ehkä jopa luovemmin.

## LÄHTEET


Anderson, Alison. 2009. Media, politics and climate change: Towards a new research agenda. *Sociology Compass* 3:2, 166–182.

Babones, Salvatore. 2016. Interpretive quantitative methods for the social sciences. *Sociology* 50:3, 453–469.

Bail, Christopher. 2014. The cultural environment: Measuring culture with big data. Theory and Society 43:3, 465–482.

Benford, Robert D. ja Snow, David A. 2000. Framing processes and social movements: An overview and assessment. *Annu. Rev. Sociol.* 26, 611–639.

Blei, David M., Ng, Andrew Y. ja Jordan, Michael I. 2003. Latent Dirichlet Allocation. *Journal of Machine Learning Research 3*, 993–1022.

Boykoff, Maxwell T. 2011. *Who speaks for the climate? Making sense of media reporting on climate change*. Cambridge: Cambridge University Press.

Boykoff, Maxwell T. ja Crow, Deserai A. 2014. *Culture, politics and climate change*. Lontoo: Routledge.

Broadbent, Jeffrey ym. 2016. Conflicting climate change frames in a global field of media discourse. Socius: *Sociological Research for a Dynamic World*, DOI: https://doi.org/10.1177/2378023116670660.

Chang, Jonathan, Boyd-Graber, Jordan, Gerrish, Sean, Wang, Chong ja Blei, David M. 2009. Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems* 22: 288–296.

DiMaggio, Paul, Nag, Manish ja Blei, David M. 2013. Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. Government arts funding. *Poetics* 41:6, 570–606.

Dunlap, Riley E. ja McCright, Aaron M. 2015. Countering climate change: the denial countermovement. Teoksessa Riley E. Dunlap ja Robert J. Brulle (toim.), *Climate change and society: Sociological perspectives*. Oxford: Oxford University Press.

Entman, Robert M. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43:4, 51–58.

Evans, Michael S. 2014. A computational approach to qualitative analysis in large textual datasets. *PLoS ONE* 9:2, 1–10.

Farrell, Justin. 2015. Network structure and influence of the climate change counter-movement. *Nature Climate Change* 6:4, 370–374.

Farrell, Justin. 2016. Corporate funding and ideological polarization about climate change. *Proceedings of the National Academy of Sciences* 113:1, 92–97.

Glaser, Barney G. ja Strauss, Anselm L. 1967. *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.

Goffman, Erving. 1974. *Frame analysis*. New York: Harper & Row.

Grimmer, Justin ja Stewart, Brandon M. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21:3, 267–297.

Hajer, Maarten. 1995. *The politics of environmental discourse: Ecological modernization and the policy process*. Oxford: Oxford University Press.

Hoffman, Andrew J. 2011. The growing climate divide. *Nature Climate Change* 1:4, 195–196.

Hoffman, Andrew J. 2015. *How culture shapes the climate change debate*. Stanford: Stanford University Press.

Hulme, Mike. 2009. *Why we disagree about climate change: Understanding controversy, inaction and opportunity*. Cambridge: Cambridge University Press.

Kukkonen, Anna, Ylä-Anttila, Tuomas, Swarnakar, Pradip, Broadbent, Jeffrey, Lahsen, Myanna ja Stoddart, Mark C.J. 2018. International organizations, advocacy coalitions, and domestication of global norms: Debates on climate change in Canada, the US, Brazil, and India. *Environmental Science and Policy* 81, 54–62.

Lamont, Michèle ja Thévenot, Laurent (toim.). 2000. *Rethinking comparative cultural sociology: Repertoires of evaluation in France and the United States*. Cambridge: Cambridge University Press.

Levy, Karen E. C. ja Franklin, Michael. 2013. Driving regulation: Using topic models to examine political contention in the U.S. trucking industry. *Social Science Computer Review* 32:2, 182–194.

McCright, Aaron M. ja Dunlap, Riley E. 2003. Defeating Kyoto: The conservative movement's impact on U.S. climate change policy. *Social Problems* 50:3, 348–373.

Moretti, Franco. 2013. *Distant reading*. Lontoo: Verso.

Nisbet, Matthew C. 2009. Communicating climate change: Why frames matter for public engagement. *Environment: Science and Policy for Sustainable Development* 51:2, 12–23.

Oreskes, Naomi ja Conway, Erik M. 2010. *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. New York: Bloomsbury.

Painter, James ja Ashe, Teresa. 2012. Cross-national comparison of the presence of climate scepticism in the print media in six countries, 2007–2010. *Environmental Research Letters* 7(4).

Purhonen, Semi ja Toikka, Arho. 2016. "Big datan" haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät. *Sosiologia* 53:1, 6–27.

Rabe, Barry G. 2010. The aversion to direct cost imposition: Selecting climate policy tools in the United States. *Governance* 23:4, 583–608.

Schmidt, Benjamin M. 2012. Words alone: Dismantling topic models in the humanities. *Journal of digital humanities* 2:1. http://journalofdigitalhumanities.org/2-1/words-alone-by-benjamin-m-schmidt/

Schmidt, Andreas ja Schäfer, Mike S. 2015. Constructions of climate justice in German, Indian and US media. *Climate Change* 133:3, 535–549.

Schäfer, Mike S. ja Schlichting, Inga. 2014. Media representations of climate change: A meta-analysis of the research field. *Environmental Communication: A Journal of Nature and Culture* 8:2, 142–160.

Timmermans, Stefan ja Tavory, Iddo. 2012. Theory construction in qualitative research: From grounded theory to abductive analysis. *Sociological Theory* 30:3, 167–186.

Trumbo, Craig W. ja Shanahan, James. 2000. Social research on climate change: Where we have been, where we are, and where we might go. *Public Understanding of Science* 9:3, 199–204.

Ylä-Anttila, Tuomas ja Kukkonen, Anna. 2014. How arguments are justified in the media debate on climate change in the USA and France. *Int. J. Innovation and Sustainable Development* 8:4, 394–408.

Ylä-Anttila, Tuomas, Vesa, Juho, Eranti, Veikko, Kukkonen, Anna, Lehtimäki, Tomi, Lonkila, Markku ja Luhtakallio, Eeva. 2018. Up with ecology, down with economy? The institutionalization of the idea of climate change mitigation in the global public sphere. Hyväksytty julkaistavaksi, *European Journal of Communication.*

Ylä-Anttila, Tuukka. 2018. Populist knowledge: 'Post-truth' repertoires of contesting epistemic authorities. *European Journal of Cultural and Political Sociology*, OnlineFirst, 9.1.2018.