Eduskunnan täysistunnon puheenaiheet 1999–2014: miten käsitellä LDA-aihemalleja?

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

PETRI LOUKASMÄKI
KIMMO MAKKONEN

ABSTRAKTI Latent Dirichlet Allocation (LDA) on yksi käytetyimmistä laskennallisista tekoälypohjaisista metodeista, joita kutsutaan aihemalleiksi (*topic models*). Esitämme ja analysoimme eduskunnan täysistuntokeskusteluista laskemiamme LDA-malleja ja arvioimme, mikä aihemäärä olisi mielekäs puheiden sisällön eksploratiiviseen analyysiin. Uutena metodisena sovelluksena analysoimme LDA-aiheiden samanaikaista esiintymistä eri puheissa osuuskorrelaatiokertoimilla. Niiden avulla LDA-mallin tuottamia aiheita voidaan käsitellä samaan tapaan kuin kahdessa saman kaltaisessa metodissa, dynaamisissa aihemalleissa ja korreloituneissa aihemalleissa (*correlated topic models* ja *dynamic topic models*), kun aineistoon sisältyy tieto puheiden ajankohdasta ja voidaan olettaa, että sanasto on pysynyt pääpiirteissään muuttumattomana analysoitavalla ajanjaksolla. Lisäksi esitämme luokituksen, jonka avulla ihmisarvioijat voivat analysoida LDA:n tuottamien aiheiden laatua. Tapausesimerkkinä esitämme korrelaatioanalyysin kuntien ja valtion suhdetta käsittelevän aiheen sekä demokratia-aiheen ja budjetti-aiheen yhteyksistä. Täysistuntokeskusteluissa on havaittavissa huomattava muutos ennen ja jälkeen vuoden 2011 eduskuntavaaleja: aiemmin keskustelu käsitteli rahaa ja valtionosuuksia, kun taas Kataisen hallituksen kuntauudistushankkeen aikana debatti käsitteli demokratiaa.

JOHDANTO

Mistä aiheista eduskunta puhuu täysistunnoissa? Kuinka täysistuntopuheenaiheiden analyysi on toteutettavissa laskennalliseen tekoälypohjaiseen tutkimusmetodiikkaan kuuluvalla tietyllä menetelmällä, eli käyttäen Latent Dirichlet Allocation (LDA) -aihemalleja (*topic models*)[1]? Aihemallit ovat bayesilaisia tilastollisia metodeja, joiden tarkoituksena on erottaa merkityksellisiksi tulkittavia luokkia jäsentelemättömästä tekstistä. Mallien etuna on mahdollisuus käsitellä kaikki eduskuntapuheet kokonaisena korpuksena, kun perinteisissä ihmislukijan analyysitavoissa jouduttaisiin tyytymään osajoukkoon puheista.

Tämän artikkelin keskeinen sanoma on LDA-metodiikassa. Analysoimme täysistuntopuheenaiheita ja kehitämme luokituksen, jolla ihmislukijat voivat arvioida LDA-mallien toimivuutta. Uutena metodisena sovelluksena analysoimme LDA-aiheiden välisiä yhteyksiä osuuskorrelaatiokertoimilla (korrelaatio aineistossa, joka summautuu vakiolukuun; eri kuin osittaiskorrelaatio), joilla LDA-aihemallia voi tarkastella samaan tapaan kuin korreloitunutta aihemallia (*correlated topic model*: Blei ja Lafferty 2007). Esimerkkitapauksena tarkastelemme kunta-puheenaiheen yhteyksiä budjetti-, demokratia- ja sote-aiheisiin. Kuntien ja valtion suhteesta puhuttiin rahoitussuhteena, kunnes vuodesta 2011 alkaen keskustelu muuttui puheeksi demokratiasta, itsehallinnosta ja poliittisen järjestelmän rakenteesta.

Analysoimme, miten hallituksen ja opposition aiheet eroavat toisistaan koko korpuksen tasolla tarkasteltuna sekä tarkastelemme kuinka LDA-mallin laadukkuutta voi arvioida kokonaisuutena sekä yksittäisen LDA-aiheen tasolla. Haluamme lisäksi suomentaa käsitteistöä aihemallinnusmetodiikalle sekä automaattiselle sisällönanalyysille yleisemminkin, sillä jonkinlainen tietämys teemasta tullee kuulumaan politiikan tutkijan perussivistykseen jo lähitulevaisuudessa (Ahonen ja Wiberg 2018).

LDA on ohjaamaton aineistolähtöinen malli, joka perustuu koneoppimiseen. Sen avulla voidaan esimerkiksi löytää eduskunnan suuresta puheiden määrästä jatkotarkasteluun tekstejä, joita ei välttämättä huomattaisi esimerkiksi tutkimalla vain tiettyjen lakiehdotusten yhteydessä käytyjä keskusteluja. Ohjatuilla aihemalleilla puolestaan voidaan muodostaa luokituksia käyttäen mallin opettamiseen referenssitekstejä, joiden positiot oletetaan tunnetuksi. Aihemallit ovat joukko tutkimusvälineitä, jotka kasvattavat ihmislukijoiden mahdollisuuksia hallita suuria aineistoja (Grimmer ja Stewart 2013) ja jotka rikastuttavat aineistolähtöistä analyysia tuottamalla uusia luokitustapoja (Günther ja Quandt 2016).

Tietokone voi monella tapaa ”lukea” tekstiä paremmin kuin ihminen sikäli, että se ei unohda lukemaansa tai valikoi vain miellyttäviä osia siitä eikä puhu lukeminaan teksteistä, joista on kuullut muualta (ks. esim. Bayard 2009). Toisaalta aihemallien tuottamien aiheiden tulkinta jää tutkijoiden vastuulle. Tavoitteenamme on lisäksi esitellä laskennallisen analyysin eduskuntatutkimukselle tuottamia uusia mahdollisuuksia (vrt. Kontula 2017).

Täysistuntokeskustelut ovat tärkeä tutkimuskohde, koska istunnoissa edustajat, eduskuntaryhmät ja ministerit perustelevat tekemiään päätöksiä ja valintoja sekä tuovat keskustelun kohteeksi ja politisoinnin piiriin tärkeinä pitämiään asioita. Parlamentti edustaa kaikkia kansalaisia – myös niitä, joilla ei ole äänioikeutta – ja parlamentti puhuu edustettavien puolesta sielläkin, missä näillä ei ole suoraa äänivaltaa, eli toimii poliittisen kontrollin välineenä kansalaisten puolesta hallituksen ja hallinnon suuntaan (Palonen 2012).

Täysistuntokeskusteluja voidaan pitää suurelta osin rinnakkaisina monologeina, mutta toisinaan käydään debattia, mitä osoittavat edustajien viittaukset toistensa puheenvuoroihin sekä esitetyt välihuudot (Pekonen 2011). Debatissakin edustajat tekevät valinnan osallistumisestaan, ja siten myös debattipuheenvuoroissa ilmenee, mitkä aiheet ja sanavalinnat ovat kullekin edustajalle ominaisia – olettaen että edustaja on saanut puheenvuoron. Kansanedustajat pääsevät puhumaan pöytäkirjaan lakiesityksiä käsiteltäessä jossain vaiheessa keskustelua, mutta suulliset kysymykset ja usein myös debattipuheenvuorot ovat niukkoja ja haluttuja, ja puhemiehet ratkaisevat, ketkä pääsevät ääneen.

Suomen eduskunnan täysistuntokeskusteluja ei ole aiemmin mallinnettu LDA:lla, eikä muilla vastaavilla aihemalleilla, toisin kuin esimerkiksi Euroopan parlamentin tai Yhdysvaltain kongressin puheita (Quinn ym. 2010; Greene ja Cross 2017; katsaus Jelodar ym. 2019). Curran ym. (2018) analysoivat Uuden-Seelannin parlamentin keskusteluja LDA:lla 40 000 000 sanan pöytäkirja-aineistolla, joka on vastaavan kokoinen kuin tässä tutkimuksessa.

Sakamoto ja Takikawa (2017) analysoivat LDA-aiheita käyttäen Japanin parlamentin ja Yhdysvaltain kongressin keskustelujen polarisoituneisuutta erilaisista teemoista puhuttaessa. Brier ja muut kirjoittajat (2016) käyttivät LDA:ta puolueposition oidenanalyysiin Italian budjettikeskusteluista. Kleynhans (2014) käytti LDA:ta Etelä-Afrikan parlamentin puheenaiheiden analyysiin, mutta hänen ensisijaisena päämääränään oli puheentunnistusmetodin kehittäminen. Yksi automatisoidun kvantitatiivisen tekstianalyysin yleistymistä politiikan tutkimuksessa hidastanut tekijä onkin tutkimustulosten julkaiseminen muiden alojen kuten kieliteknologian ja tietojenkäsittelytieteen foorumeilla (Boumans ja Trilling 2016).

Kotimaisessa yhteiskuntatieteellisessä tutkimuksessa ensimmäisenä vertaisarvioituna julkaisuna LDA:ta hyödynsivät Purhonen ja Toikka (2016), jotka analysoivat havainnollistuksena tasavallan presidenttien kaikki uudenvuoden puheet. Politiikan tutkimuksessa LDA-mallia ovat soveltaneet Winter ja Wiberg (2016) analysoidessaan vuoden 2015 aikana Suomessa säädettyjä lakeja, mutta he eivät raportoi mallinsa toimivuuden diagnostiikkaa. Liu ja Jansson (2017) puolestaan tutkivat LDA:lla Helsingin seudun kaupunkitapahtumiin liittyvää viestintää Instagramissa. Laaksonen ja Nelimarkka (2018) analysoivat vaalijulkisuutta ja aiheomistajuutta vuoden 2015 eduskuntavaaleissa, mikä muodostaa kiintoisan teoriataustan myös täysistuntopuheiden analyysille, koska eduskunnassa lausuttua voi pitää myös viestintänä kansalaisille tulevia vaaleja ajatellen. Nelimarkka (2019) analysoi suomalaisia puolueohjelmia esimerkkinä käyttäen aihemallinnuksen soveltamista yhteiskunnallisiin teksteihin ja tulosten yhteyttä politologiseen teoriaan sekä tilastollisten menetelmien hyödyntämistä parametrivalinnassa niin, että toimitaan aidosti aineistolähtöisesti.

LDA-mallin laadun arviointi vaatii paljon työtä, ja sitä varten on kehitetty laskennallisia apuvälineitä ja visualisointityökaluja. Kehitimme luokituksen, jonka avulla ihmislukija voi arvioida mallien toimivuutta, vertailla malleja keskenään sekä ratkaista, mitä LDA:n tuottamia aiheita jatkoanalyysissa on mielekästä käyttää. Luokituksemme eroaa aiemmista ihmislukijoille laaditusta analyysikehikoista siten, että se soveltuu hyvin aineistolähtöiseen tutkimusotteeseen, kun muissa luokituksissa tavoitteena on ollut löytää mahdollisimman hyvä yhteys ennalta tunnettuun luokitukseen tai valmiiksi luokiteltuun vertailuaineistoon.

Esityksen yksinkertaistamiseksi käsittelemme eduskuntaryhmiä yhtenäisinä toimijoina ikään kuin kukin ryhmä olisi yksi puhuja. Aihemalleilla vertailua voi tehdä myös puhujatasolla ja näin

saada esille ryhmien sisäisiä eroja, jotka jäävät usein äänestyksissä näkymättömiin ryhmäkurin paineessa. Esimerkiksi Curran ym. (2018) loivat LDA-aiheiden perusteella verkostomallin, jonka avulla on havainnollistettavissa edustajatasolla mistä aiheista muodostuu yhtenäisiä puhujajoukkoja ja mistä koko parlamentti puhui esimerkiksi yllättävien ulkoisten tapahtumien seurauksena.

Esimerkkitapauksena kuvailemme kuntien asemaan ja kunta–valtio-suhteeseen liittyvän keskustelun esiintymistä täysistunnoissa. Päämääränämme on kuvata suomenkielisen aineiston LDA-mallintamiseen liittyvät haasteet sekä ratkaisut, joiden avulla eduskunnan puheenaiheet saadaan mallinnettua mielekkäällä tavalla. Kun malli on laskettu ja havaittu laadukkaaksi, sitä voidaan käyttää mm. eri aiheiden keskinäisten yhteyksien analysointiin. Esimerkiksi puhuttaessa kuntien asemasta voidaan selvittää, ketkä puhuvat samassa yhteydessä rahasta, demokratiasta, asukkaiden oikeuksista tai muista teemoista. Vaikka eri aiheiden yhteyksien analyysiin ei mentäisi, jo tieto aiheista on sinänsä usein kiinnostavaa. Esimerkiksi onko nais- ja miesedustajilla erilaiset puheenaiheet? Miten aiheiden jakauma on muuttunut ajan kuluessa? Artikkelin keskeisin päämäärä on kuitenkin metodiikan kehittäminen yhdistämällä LDA ja korrelaatioanalyysi[2].

MIKÄ LDA OIKEIN ON?

LDA on ohjaamaton aihemalli, jossa analyysiyksikkönä on yksi teksti (tai muu diskreetti aineisto). Tutkija määrittelee, kuinka moneen aiheeseen LDA luokittelee tekstit. LDA tekee niin antaen kullekin tekstille todennäköisyyden, jolla teksti kuuluu kuhunkin aiheeseen, eli LDA on moniaihemalli (*mixed membership*). (Blei ym. 2003.) Vastaavasti yksiaihemallit asettavat kunkin tekstin tasan yhteen aiheeseen (esim. Greene ja Cross 2017). LDA-mallin logiikka toimii kuitenkin toisinpäin kuin klusteroinnissa tapahtuisi: jokaisen aiheen katsotaan generoivan tekstin todennäköisyydellä p, ja yhdessä kaikki aiheet generoivat koko korpuksen tekstit, joskin käytännössä mallin tuloksia voi tulkita kuten klusteroinnissakin. Dokumenttien sisältämien sanojen perusteella mallinnetaan samanaikaisesti sekä LDA-aiheiden jakauma kussakin dokumentissa että sanojen jakauma kussakin aiheessa.

Aihemallien yhteydessä aihe (*topic*) on formaali käsite, ja sanan arkikielisempään merkitykseen puheena olevasta asiasta viittaamme tässä artikkelissa sanalla teema. Muodollisesti aihe on sanaston todennäköisyysjakauma, eli yksittäisessä aiheessa jokainen sana saa todennäköisyyden, jolla se kuuluu aiheeseen (Blei 2012, 78). Aiheet ovat LDA:ssa eräänlaisia abstrakteja, latentteja ulottuvuuksia, joiden perusteella puhe muodostuu. Esimerkiksi arkiymmärryksessä voidaan ajatella, että puhujaa kiinnostaa perhepolitiikka tai puolustuspolitiikka, ja nämä latentit ulottuvuudet generoivat yksittäisen puheen sisällön, joka ilmenee perhe- tai puolustussanastona.

Moniaihemallit soveltuvat eduskunnan täysistuntokeskusteluiden analyysiin, koska edustajat käsittelevät puheissaan erilaisia aiheita ja näkökulmia sekä toisinaan päätyvät sivupoluille. Toisaalta Greene ja Cross (2017) pitävät yksiaihemalleja toimivina Euroopan parlamentin puheiden analyysiin, koska puheenvuorojen tiukka aikaraja pitää parlamentin jäsenet yhdessä puheenaiheessa.

Sanojen ja tekstien tiettyyn LDA-aiheeseen kuulumisen todennäköisyys estimoidaan ennen mallin laskentaa asetettujen parametrien puitteissa. Parametreista tärkein on aiheiden määrä.

Pieni aihemäärä tuottaa yleensä hyvin laajoja teemoja, ja suuri aihemäärä johtaa kielen ominaisuuksien toistumiseen. Esimerkiksi sanat ’ministeri’ ja ’kysyä’[3] saattavat mallintua omaksi aiheekseen, mikäli aihemäärä on suuri. Lopputulokseen vaikuttaa myös jäljempänä esittelemämme aineiston esikäsittely, kuten käytetäänkö jäsenninohjelmia sanojen muuntamiseksi perusmuotoonsa ja poistetaanko erittäin yleiset tai erittäin harvinaiset sanat aineistosta.

Tässä yhteydessä emme esittele LDA-mallin formaalia määritelmää, joka on kuvattu tarkasti muualla tutkimuskirjallisuudessa (ks. esim. Blei ym. 2003 tai yleistajuisemmin Blei 2012). LDA-mallinnuksessa havaittu informaatio on W sanaa, jotka esiintyvät M kappaleessa dokumentteja. Muilta osin kyse on latenteista muuttujista (aiheiden jakauma kussakin dokumentissa θ_m ja sanojen jakauma kussakin aiheessa ϕ_k), joiden arvot estimoidaan mallinnuksessa.

LDA perustuu sanakorioletukseen (*bag of words*): kustakin puheesta analysoitava informaatio on puheen sanojen frekvenssi. Sanojen esiintymisjärjestyksellä ei ole merkitystä. Käytännössä kyse on siitä, että toimiessaan toivottavalla tavalla LDA antaa yksittäisessä aiheessa suuren todennäköisyyden sanoille, joita käytetään samasta teemasta puhuttaessa. Useimmiten LDA:n tuottamille aiheille löytyy myös vastine arkikielessä. Esimerkiksi täysistuntopuheita, joissa esiintyy tiuhaan sanat ’lapsi’, ’perhe’, ’vanhempi’, ’nuori’, ’oikeus’, ’äiti’ ja ’isä’, voidaan pitää teemaltaan perhepolitiikkaan kuuluvina. Toisaalta nuorista ja oikeuksista puhutaan usein muissakin asiayhteyksissä, kuten koulutuspolitiikkaan tai syrjäytymiseen liittyvissä teemoissa, jotka LDA mallintaa yleensä omiksi aiheikseen täysistuntopuheita analysoitaessa.

LDA:ta käytetään myös yksiaihemallin tapaan raportoimalla merkitsevin aihe (esim. Ylä-Anttila ym. 2018), mutta silloin jätetään hyödyntämättä informaatiota aiheiden osuuksista samaan tapaan kuin jos puolueiden valtaa Suomen kunnissa analysoitaisiin raportoimalla kunkin kunnan suurin puolue. Yksinkertaisten sekä hyvien moniaihemallien raportointi- ja visualisointikeinojen kehittämiselle on tarvetta samaan tapaan kuin puolueiden valtasuhteita voidaan tarkastella efektiivisten puolueiden lukumäärän tai valtaindeksien avulla.

LDA-mallilla ei ole yksiselitteistä matemaattista ratkaisua. Siksi jokainen laskentakerta tuottaa hieman erilaisen ratkaisun samalle aineistolle, vaikka aihemäärä ja hyperparametrit[4] säilytettäisiin muuttamattomina. Kun malli on laskettavissa monia kertoja, aiheiden vakauden (stabiliteetti) ja erottelukyvyn tarkastelu tulee tarpeelliseksi: mitkä aiheet toistuvat laskennasta toiseen, ja sopiiko jokin malli erityisen hyvin jonkin tietyn kysymyksen analyysiin. Esimerkiksi saamelaisteema erottuu joissain malleissa omassa aiheessaan, kun taas toisilla laskentakerroilla teema voi esiintyä Pohjois-Suomeen, vähemmistöoikeuksiin ja kalastukseen liittyvien aiheiden yhteydessä[5].

LDA määrittelee jokaisen sanan todennäköisyyden jokaisessa aiheessa, ja yksittäinen sana saattaa esiintyä suurehkolla todennäköisyydellä monissa aiheissa. Eduskunta-aineistossa tällaisia sanoja ovat yleensä esimerkiksi ’Suomi’, ’hallitus’ ja ’euro’, joiden esiintymisen perusteella on vaikeaa kuvailla, mihin teemaan aihe liittyy. Ihmislukijalle helpommin tunnistettavia ovat erityisalasanat, ja esimerkiksi kun edellisten yhteydessä suuren todennäköisyyden saavat mm. ’määräraha’, ’valtiovarainvaliokunta’ ja ’talousarvio’, aihetta voi kuvata budjettiteemaiseksi. Lopulta aiheiden määrä ratkaisee, kuinka pieniä yksityiskohtia malli kykenee erottelemaan. Toisaalta, jos aiheita on satoja tai tuhansia, LDA-aiheet erottelevat isolta osin kielelle ominaisia rakenteita ja sanontoja, eikä aiheille ole aina löydettävissä yhdistävää teemaa samaan tapaan kuin pienemmillä aihemäärillä. Nelimarkka (2019) ottaa esille LDA-aiheiden vastinpareiksi yhteiskuntatieteellisessä teoriassa muun muassa kehykset (*frames*), teemat (*theme*) ja asiat (*issues*).

Pienillä aihemäärillä erottunee laajempia kehyksiä ja suuremmilla rajatumpia asioita: esimerkiksi kokeilimme 20-aiheista mallia, jossa erottui sosiaalipolitiikka-kehys, ja useammilla aihemäärillä erilaisia sosiaaliturvan muotoja.

Paras aihemäärä riippuu tutkimuskysymyksestä, ja esimerkiksi avoimien kyselyvastauksien erottelussa se lienee pienempi kuin parlamentin puhekorpuksessa. Mielekkäiden aihemäärien haarukointiin on useita mittalukuja (ks. esim. R-paketti *ldatuning*), joista tavallisimmat ovat perpleksiteetti ja logaritmisen uskottavuuden arvo (Nelimarkka 2019, 11–12). Eduskunta-aineistossamme perpleksiteettiarvot viittasivat alustavissa tutkimuksissa suurten aihemäärien käyttöön, mutta silloin aiheiden tulkinta ja raportointi olisi ollut hankalaa. Sen sijaan päädyimme haarukoimaan erilaisia aiemmissa tutkimuksissa esiintyneitä aihemääriä, vaikkei lähestymistapaa voikaan silloin pitää puhtaasti aineistolähtöisenä.

AINEISTO JA SEN ESIKÄSITTELY

Aineistomme koostuu 181 539 täysistuntopuheesta valtiopäivävuosilta 1999–2014, joista valtiopäivävuodelta 1999 aineistoon sisältyvät vain pöytäkirjat 86–117, ja muut vuodet kokonaisuudessaan. Valtiopäivien avajaisia ja avajaisjumalanpalveluksia ei ole huomioitu ja puhemiespuheenvuorot on poistettu[6]. Lukuisat puheet (2 844 kpl) sisälsivät sekä suomen- että ruotsinkielisiä osia, joista analyysiin on sisällytetty suomenkieliset kappaleet. Monikielisten aineistojen analyysiin on kehitetty algoritmeja, mutta yksinkertaisuuden vuoksi keskitymme tarkastelemaan suomenkielisiä puheita, joita pidettiin jokaisessa eduskuntaryhmässä. Ruotsinkielisten puheiden aihejakauma eroaa suomenkielisistä, ja esimerkiksi Ahvenanmaan asema, vähemmistöoikeudet ja kielikysymykset erottuvat niissä useammin omina aiheinaan.

Valtiopäivävuodesta 2015 alkaen eduskunta muutti julkaisujärjestelmäänsä, minkä vuoksi uusimmat täysistuntopuheet eivät sisälly tähän korpukseen. Uusien puheiden yhdistäminen aineistoomme vaatisi laajan tietojenkäsittelytyön, mikä on mielekästä toteuttaa vasta, kun eduskunnan avoin internetrajapinta valmistuu ja vakiintuu. Analysoimme pöytäkirjoja olettaen niiden vastaavan täysin puhujan lausumaa. Vaikka täysistuntopöytäkirjojen tekstiä toimitetaankin, puheiden leksikaalisia valintoja ei liiemmin muuteta (Voutilainen 2016, 179), mikä tekee niistä erittäin käyttökelpoisia sanastoon perustuvalle analyysille. Pöytäkirjatekstin kieliasun muokkaaminen kielenhuoltonormeja vastaavaksi helpottaa aineiston konelukemista, ja sen ansiosta aihemallinnuksessa käytettävä informaatio saadaan kerättyä kattavasti.

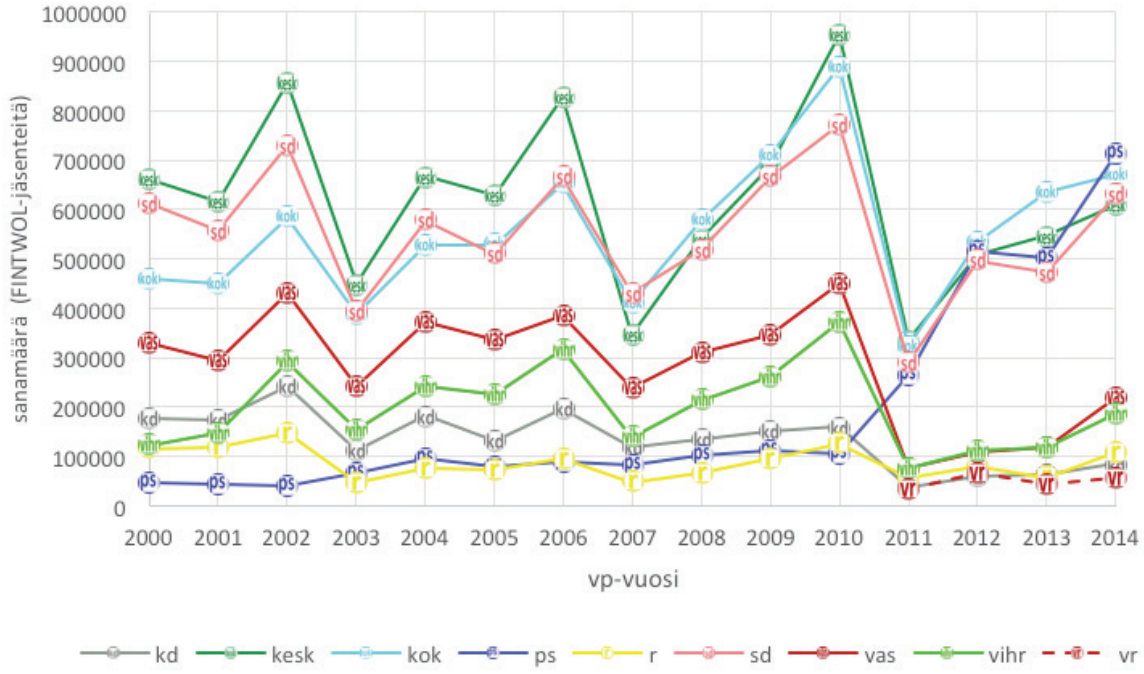
Aineisto jäsennettiin Lingsoft Oy:n FINTWOL-jäsentimellä sekä Turku BioNLP -ryhmän Finnish Dependency Parserilla (FDP)[7]. FINTWOL perustuu rajoitekielioppiin ja FDP riippuvuuskielioppiin, minkä vuoksi ne tuottavat hieman erilaisia sanamääriä, vaikka molemmat tunnistavat suomen sanaston hyvin. FINTWOL muunsi sanat perusmuotoonsa, tai mikäli mahdollisia perusmuotoja oli useita, esitti jokaisen. Esimerkiksi tekstissä esiintyvä sana ’teille’ voi olla peräisin sanoista ’te’ tai ’tie’, jolloin jäsennin palautti molemmat vaihtoehdot putki-merkillä erotettuna: ’te|tie’. Näin korpuksesta muodostui 76 097 erilaista jäsennettä[8], joita esiintyi yhteensä 40 304 399 kappaletta. FDP puolestaan palautti sanan perusmuodon, jos jäsennin tunnisti sanan. FDP tuotti 68 881 jäsennettä, joita esiintyi 39 519 258 kappaletta, kun välimerkkejä ja numeraaleja ei huomioida. Sanat, joita jäsentimet eivät tunnistaneet, jätettiin huomiotta.

Jäsenteistä poistettiin sulkusanat eli kaikkein yleisimmin esiintyvät jäsenteet, kuten pronominit ja tavallisimmat partikkelit sekä numerot, päivämäärät ja yhden merkin mittaiset sanat, poikkeuksena ’§’. Täysistuntokeskusteluiden erikoisuutena sulkusanoihin kuuluivat myös ’arvoisa’, ’rouva’, ’herra’ ja ’puhemies’, jotka esiintyvät likimain jokaisessa puheenvuorossa. Lisäksi sanastosta poistettiin jäsenteet, jotka esiintyivät korpuksessa alle viisi kertaa. Sulkusanojen poisto on ensisijaisesti laskentatekninen kysymys, koska siten aineistoa saadaan pienennettyä. Suurten sulkusanalistojen laadinta ei kuitenkaan kohenna tulosta merkittävästi, sillä malli huomioi yleisesti esiintyvät sanat hyvin (Schofield ym. 2017). LDA muodostaa yleissana-aiheita, joihin mallintuvat kaikille puheille ominainen sanasto, kuten ’käydä’, ’haluta’, ’ihminen’, ’silloin’, ’hyvin’ yms., ja tavallisesti erityisalasanasto mallintuu omiksi aiheikseen.

Puheinformaation määrässä on suuria eroja eduskuntaryhmittäin ja edustajittain. Edustajat saavat puheenvuoron ilmoittautumisjärjestyksessä (Eduskunnan työjärjestys 1999/2000, 50 §), ja edustajilla pitäisi siten olla mahdollisuus päästä puhumaan. Suulliset kyselytunnit ovat tästä poikkeus, kun puheaikaa riittää vain niille, joille puhemies sitä suo, eikä debattipuheenvuorojakaan välttämättä riitä kaikille halukkaille. Voutilainen (2016) erottelee täysistuntokeskustelujen genreiksi edellisten lisäksi keskustelun lakiesityksestä tai -aloitteesta, budjetista, valtioneuvoston selonteosta tai tiedonannosta, pääministerin ilmoituksesta tai välikysymyksestä sekä ajankohtaiskeskustelun. Niissäkin puheenvuorotyypit vaihtelevat, kuten ministerin tai valiokuntapuheenjohtajan esittelypuheenvuoro, ryhmäpuheenvuoro, edustajan ”varsinainen” puheenvuoro, lyhyt vastauspuheenvuoro sekä välihuuto salista.

Aineistossamme kaikki puheenvuorot on analysoitu samalla tavoin, vaikka esimerkiksi ryhmäpuheenvuorot ja ennen internetlähetysten yleistymistä tv-aikana pidetyt puheet ovat yleensä tärkeämpiä kuin ns. puhuminen pöytäkirjaan, jota runsaissa määrin tehneet ovat saattaneet saada ”tilastopuhuja”-liikanimen. Tältä osin edustajien toiminta lienee muuttunut tutkimusajanjaksolla, kun valtiopäivätoimille voi saada enenevässä määrin huomiota internetissä, mutta asian täsmällisempi tarkastelu rajautuu tämän artikkelin ulkopuolelle.

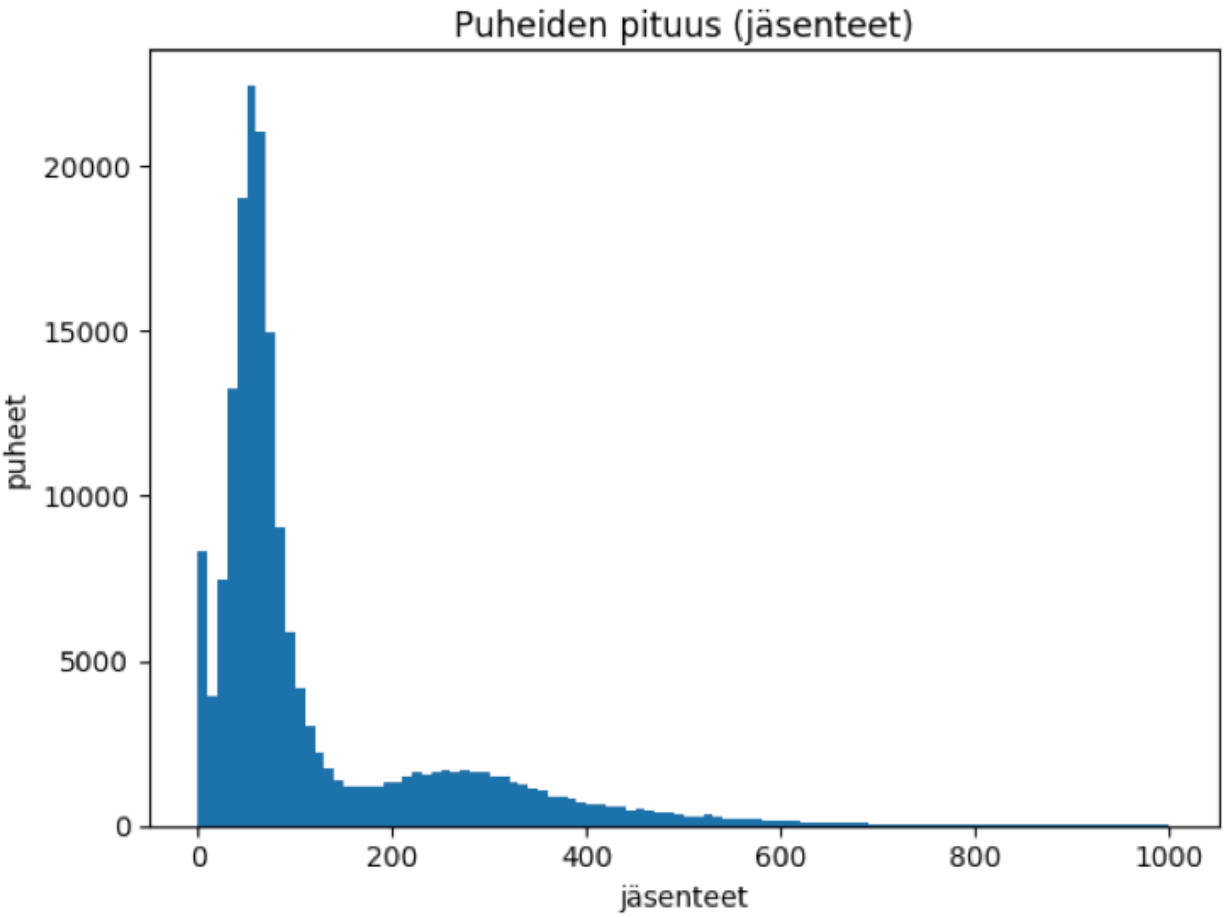
Suurissa ryhmissä on enemmän edustajia, ja siten yleensä enemmän puhetta. Kuviosta 1 ilmenee suuret erot sanamäärissä eri eduskuntaryhmillä sekä se, että jokainen koko tarkastelujakson ajan toiminut ryhmä on puhunut niin paljon, että ryhmätason analyysiin aineistoa on riittämiin. Korpuksessa eniten sanoja on keskustalta, 23,4 prosenttia. Kokoomuksen osuus on 21,2 %, sosiaalidemokraattien 21,1 %, vasemmistoliiton 10,8 %, vihreiden 7,5 % ja perussuomalaisten 7,2 %, vaikka puolueen puhemäärä kasvoikin vuoden 2011 vaalien jälkeen. Pienryhmistä vasenryhmä puhui kokoonsa nähden ahkerasti, ja se muodostaa 0,5 prosenttia suomenkielisestä kokonaissanamäärästä.



Kuvio 1. Puheiden sanamäärät eduskuntaryhmittäin ja valtiopäivävuosittain. Yhden edustajan ryhmiä ei ole huomioitu.

Puheiden pituuksissa on suurta vaihtelua (kuvio 2). Puheiden pituuden keskiarvo on 218,8 ja mediaani 125 FINTWOL-jäsennettä. Pisimmässä puheessa esiintyi 3 024 jäsennettä. Pituusvaihtelu on paljolti seurausta puhemiesneuvoston suosittelemista aikarajoista, kuten 15 minuuttia ryhmäpuheenvuorolle, 10 tavalliselle puheenvuorolle, 2 vastauspuheenvuorolle, 5 nopeatahtiselle keskustelulle ja 1 suullisen kyselytunnin puheenvuorolle. LDA-malliin muodostuu usein luokka erittäin lyhyille puheille, kun erityisesti talousarviokeskusteluissa pidetään paljon kannatuspuheenvuoroja. Silloin muodostuu aihe, jossa sanat ’kannattaa’ ja ’ehdotus’ saavat suuren painon.[9]

Mallissamme edustajia tai ministereitä ei ole painotettu aseman mukaan, vaikka edustajan senioriteetti sekä asema puolueen tai ryhmän johdossa vaikuttanee siihen, keiden puheita käytännössä kuunnellaan tarkasti. Tarkastelun ulkopuolelle jää myös kysymys siitä, vaikenevatko edustajat asioista, joista heillä muodollisesti olisi oikeus puhua. LDA huomioi vain sen, mitä on lausuttu. Tekstien esikäsittely on suoritettavissa mielekkäästi kummallakin tarkastelluista jäsentimistä, ja kyseinen työvaihe on huolellisuutta ja aikaa vaativa rutiinitoimi automatisoidussa sisällönanalyysissa. Tarvittavat työkalut suomenkieliseen käsittelyyn ovat saatavilla.



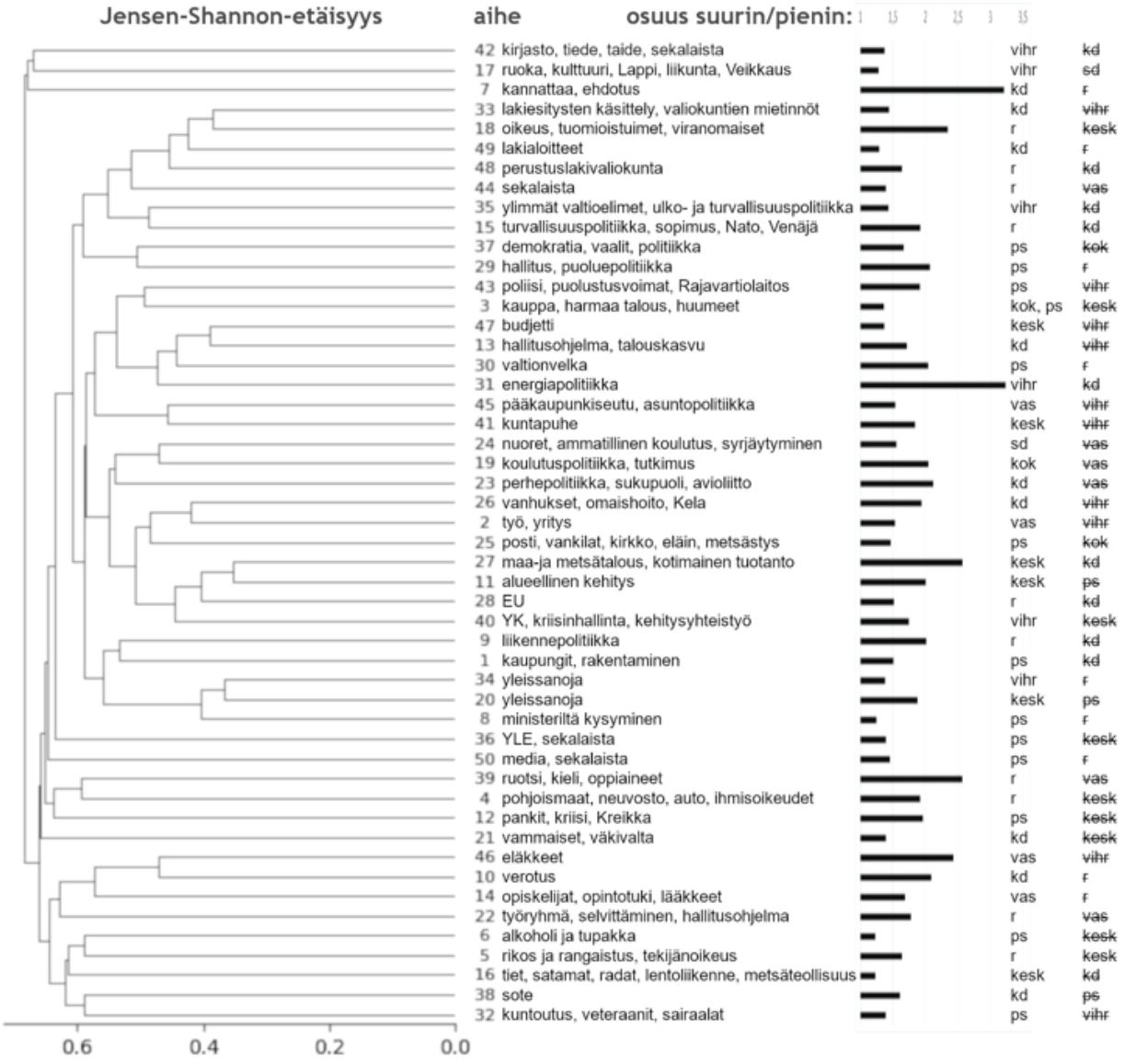
Kuvio 2. Täysistuntopuheiden pituudet, FINTWOL-jäsenteitä

TÄYSISTUNTOKESKUSTELUISTA LASKETUT LDA-MALLIT

Olemme laskeneet tutkimuksen edetessä lukuisia LDA-malleja, joiden avulla olemme arvioineet LDA:n soveltuvuutta täysistuntokeskustelujen analyysiin ja testanneet tietojenkäsittelyllisten haasteiden ratkomista. Nyt raportoimme mallijoukon, jossa olemme laatineet 30-, 50-, 60-, 70-, 80- ja 200-aiheisia malleja kaksi kappaletta kutakin (merkitty T30a, T30b, T50a jne.). Määrät valittiin siten, että niissä todennäköisesti on moniin politiikan tutkimuksen kysymyksiin sekä liian vähä- että runsasaiheisia malleja. Suurten aihemäärien ihmistulkinta on työlästä, ja siksi samoilla aihemäärillä vertailtiin vain kahta mallia. Muissa käyttötarkoituksissa, kuten seulottaessa internetin informaatiovirtaa tiedustelutehtävissä, aiheita voi olla satoja, jopa tuhansia. Käytimme FDP-jäsenteitä, koska pienempi erilaisten jäsenteiden kokonaismäärä on laskennallisesti kevyempi vaihtoehto.

Esimerkkinä LDA-mallilla löydetyistä latenteista ulottuvuuksista esittelemme 50-aiheisen mallin kuvioissa 3, 4 ja 5. Dendrogrammissa (kuvio 3) vaakasuuntainen etäisyys kuvaa aiheiden sanastojen etäisyyttä, ja tutkimusryhmä on nimennyt aiheet ihmislukijalle hahmotettaviksi. Dendrogrammi on laskettu koko aineiston perusteella, ja sen oikeaan laitaan on lisätty suhdeluku, jossa verrataan kustakin aiheesta eniten ja vähiten puhunutta ryhmää aiheiden toden-

näköisyyksien summalla koko aineistossa. Osuusanalyysissa tarkastelu rajoittuu kahdeksaan koko tarkastelujakson ajan toimineeseen eduskuntaryhmään. Vasenryhmän säilyttäminen analyysissa olisi painottanut kauden 2011–2014 ajankohtaisaiheita, ja pieni ryhmä olisi monissa aiheissa ollut eniten tai vähiten puhunut joukko.



Kuvio 3. Dendrogrammi 50-aiheisesta mallista (T50a, Jensen-Shannon-etäisyys 50 merkitsevimmästä sanasta; ihmislukijat ovat nimenneet aiheet saman sanamäärän mukaan. Oikean laidan palkit kuvaavat ko. aiheen todennäköisyyksien summien suhdetta eniten ja vähiten puhuvan ryhmän välillä: leveämpi palkki merkitsee suurempaa eroa ryhmillä. Eniten puhuva ryhmä on merkitty tavallisella tekstillä, vähiten puhuva yliviivattuna. Dendrogrammi on laskettu koko aineiston perusteella; osuusanalyysissa mukana kahdeksan koko jakson toiminutta ryhmää.

Tutkijoiden merkitsevimpien sanojen perusteella tekemien arvioiden perusteella havaittiin, että useat teemat toistuvat selkeinä ja pääsääntöisesti laadukkaina aiheina lähes kaikissa malleissa eri aihemäärillä. Tällaisia LDA-aiheissa toistuvia politiikkateemoja ovat kunta-asioiden lisäksi budjetti, energiapolitiikka, EU, koulutus, yliopistot, opiskelijat ja opintotuki, sosiaali- ja terveysasiat, työelämä, yritykset, liikenne sekä ulko- ja turvallisuuspolitiikka. Politiikka-, vaalit ja demokratia -aihe esiintyy useissa malleissa, muttei kaikissa.

Täysistuntokeskustelulle ominaisia aiheita ovat prosessiaiheet, joissa edustajat viittaavat eduskunnassa tapahtuvaan toimintaan. Tällaisia selkeästi omiksi teemoikseen mallintuvia aiheita ovat eduskuntatyö, hallitus ja oppositio sekä puoluepolitiikka, edustaja- ja puheenvuoroviittaukset, lakialoitteet, ministeriltä kysyminen ja valiokuntien mietinnöt.

Säännöllisesti esiintyviä, mutta monissa malleissa erilaisin yhdistelmin esiintyviä teemoja ovat mm. hinnat, harmaa talous, hallitusohjelma, aluepolitiikka, alkoholi, kauppa ja aukioloajat, Kela, posti, nuoret, vanhukset, eläkkeet, lapsiperheet, maatalous, maahanmuutto, ruotsi, perustuslaki, asuminen, rakentaminen ja kaupungit.

Esimerkiksi aluepolitiikka-teema esiintyy jokaisessa mallissa, mutta harvoin selkeänä omana aiheenaan. Siihen liittyviä teemoja ovat mm. Pohjois-Suomi, pääkaupunkiseutu, asuntopolitiikka ja maakunnat. Poliisi-teema puolestaan esiintyy yksinään tai yhdessä Rajavartiolaitokseen, tulliin, rikollisuuteen, vankeihin ja tuomioistuimiin liittyvän sanaston kanssa. Poliisi-teema on yleensä sikäli laadukkaammin mallintunut, että em. teemoja voi pitää hyvin läheisinä poliisitoiminnalle, kun taas aluepolitiikka-aiheessa yhteys on usein etäisempi. Vastaavasti verotuksesta puolestaan puhutaan joissain malleissa omissa aiheissaan ja toisissa verotussanasto mallintuu samaan aiheeseen eläke-, toimeentulo- ja tulonsiirtosanaston kanssa. Erillisten tai toisilleen läheisten teemojen yhdistymisen järjestelmällinen arviointi oli motivaationa jäljempänä esittämämme luokituksen kehittelylle.

Kuviossa 4 LDA-aiheiden todennäköisyyksien summat esitetään hallitus- ja oppositioryhmien mukaan eroteltuna. Koska kyseessä on todennäköisyydet, kukin puhe on aineistossa painolla yksi. Enemmistö puheista on hallitusryhmiltä (53,4 prosenttia, 96 896 kpl). Kuviosta nähdään, että hallitusryhmillä yleissana-aiheet ovat paljon yleisempiä kuin oppositiolla, joka puolestaan selvästi suuremmin edustettuna aiheissa ”ministeriltä kysyminen” ja ”hallitus, puoluepolitiikka”.

Kuvio 4. LDA-aiheiden todennäköisyyksien summat mallissa T50a hallitus- ja oppositioryhmittäin. Aiheet on järjestetty oppositio/hallitus-suhdeluvun mukaan siten, että vasemmalle sijoittuvat aiheet, joissa hallitusryhmien osuus on suurin.

Opposition puheissa aiheet ”hallitus, puoluepolitiikka”, ”verotus” ja ”hallitusohjelma, talouskasvu” esiintyvät huomattavasti suuremmassa määrin kuin hallituksella. Oppositio puhui enemmän myös aiheista ”pankit, kriisi, Kreikka”, ”demokratia, vaalit, politiikka”, ”pääkaupunkiseutu, asuntopolitiikka” sekä ”eläkkeet”. Hallitus puhui oppositiota enemmän alueellisesta kehityksestä, EU:sta, budjetista, turvallisuuspolitiikasta ja sopimuksista, maa- ja metsätaloudesta sekä perustuslakivaliokunnan lausunnoista. Koulutuspolitiikka ja tutkimus oli hallituksen aihe, kun taas opiskelijat ja opintotuki esiintyivät useammin opposition puheissa.

Useat aiheet, kuten ”rikos ja rangaistus, tekijänoikeus”, ”perhepolitiikka, sukupuoli, avioliitto”, ”kuntapuhe” sekä ”alkoholi ja tupakka”, esiintyivät melko tasaisesti hallituksella ja oppositiolla. Näissä erot tulevat esiin vasta, kun aiheita analysoidaan muulla tavoin jaoteltuna. LDA-malli tuottaa niin paljon informaatiota, ettei kaikkien aiheiden esittäminen yksityiskohtaisemmin ole mahdollista. Jäljempänä analysoimme kuntapuhe-aiheen eduskuntaryhmittäin ja valtiopäivävuosittain eroteltuna.

LDA MALLIEN LAADUN DIAGNOSTIIKKAA: C_V-KOHERENSSI

LDA-mallin laadukkuuden analysointiin on käytettävissä laskennallisia tunnuslukuja, joista käytämme C_V-vektorikoherenssia (Röder ym. 2015), sekä ihmislukijan analysoimiin otoksiin perustuvia menetelmiä. Ihmislukija voi analysoida mallia tarkastelemalla suurimpia todennäköisyyksiä kussakin aiheessa saavien sanojen joukkoa tai käyttäen visualisointityökaluja. Laskennallisten koherenssiarvojen etuna on, että niillä arvioita saadaan jatkuvalta jakaumalta, kun ihmisarvioijat tyytyvät usein paljon karkeampiin luokituksiin, esimerkiksi asteikolla yhdestä kymmeneen.

Röder ym. (2015) kehittävät puitteet erilaisten koherenssin tunnuslukujen analyysille ja muodostavat niistä yhdistelmän, joka toimii vertailukorpuksissa yhdenmukaisesti ihmislukijoiden tulkintojen kanssa. Laskimme koherenssiarvot Gensim-ohjelmakirjaston models.coherencemodel-työkalulla. C_V-koherenssi voidaan laskea halutulle määrälle todennäköisimpiä sanoja, ja jäljempänä käytämme vertailun vuoksi 10, 20, 30 ja 50 sanan koherenssilaskelmia.

Tilarajoitteen vuoksi emme tässä yhteydessä esittele C_V-koherenssin formaalia määritelmää emmekä laskenta-algoritmia. Erilaisia koherenssin laskentatapoja on lukuisia. Abstraktilla tasolla Röder ym. (2015) vertaavat, että sanastokoherenssin käsite on analoginen tietoteorian koherenssin kanssa vaihtamalla väitteen tilalle sana: samoin kuin voidaan analysoida, esiintyvätkö tosiasiaväitteet yhdenmukaisena joukkona, voidaan tarkastella, esiintyvätkö sanat yhdessä. Koherenssin laskennassa sanat jaetaan osajoukkoihin, sanajoukoille (voi olla myös yhden sanan joukko) lasketaan todennäköisyydet esiintymiselle muiden sanajoukkojen yhteydessä, ja näistä todennäköisyyksistä koostetaan yhtenäinen suure. Tätä lukua vertaamme jäljempänä ihmislukijoiden arvioihin LDA-mallien toimivuudesta. Mitä suurempi koherenssi on, oletettavasti sitä paremmin aihe on mallintunut.

LDA-MALLIN DIAGNOSTIIKKAA: IHMISLUKIJAN LUOKITUKSET

Yksinkertaisin ihmislukijan analyysitapa on tarkastella kunkin aiheen merkitsevimpiä sanoja, ja arvioida muodostuuko niistä mielekäs kokonaisuus. Usein aihetta kuvaillaan raportoimalla 10–15 sanaa, mutta analyysissa kannattaa huomioida tätä suurempia sanajoukkoja. Analysoimme aineistomme käyttäen 50 merkitsevimmän sanan joukkoa sekä siinä esiintyvien sanojen todennäköisyyksiä aiheissa (kuvio 5). Juuri 50 sanan tarkastelu on sikäli mielivaltainen valinta, että 49 tai 51 merkitsevimmän sanan analyysi tuottaisi mitä todennäköisimmin saman tuloksen. Pienemmälläkin sanamäärällä pystyisi yleensä tekemään johtopäätöksen hyvin, mutta käytimme 50 sanaa voidaksemme olla varmoja tulosten laadusta. Purhonen ja Toikka (2016, 18) käyttävät samaa 50 kärkisanaa aiheiden nimeämisessä, ja Ylä-Anttila ym. (2018) kymmentä sanaa sekä kymmentä aiheen merkitsevintä lähdetekstiä. Mitä enemmän aiheita on, sitä pienemmän sanajoukon ihmislukija-analyysi riittää, koska yksittäisissä aiheissa sanojen todennäköisyydet pienenevät nopeammin.

topic #6
documents 40928 (22.57)%, top documents 4959 (12.12%)

koulutus	0.0385	perus#opetus	0.0064	mahdollisuus	0.0053	valmistua	0.0035
nuori	0.0371	rahoitus	0.0061	aste	0.0051	järjestää	0.0033
yli#opisto	0.0279	osaaminen	0.0060	laatu	0.0051	esimerkiksi	0.0033
opiskelija	0.0224	syrjäytyä	0.0059	lukio	0.0050	vahvistaa	0.0033
koulu	0.0176	sivistys#valio#kunta	0.0059	tutkinto	0.0050	tukea	0.0033
opetus	0.0131	opiskelu	0.0059	tavoite	0.0049	huomio	0.0032
ammatillinen	0.0117	lisätä	0.0058	opiskella	0.0048	työelämä	0.0032
tärkeä	0.0114	opinto	0.0058	opinto#tuki	0.0048	erityisesti	0.0032
kehittää	0.0108	opetus#ministeriö	0.0057	suorittaa	0.0047	yhteis#työ	0.0032
opettaja	0.0105	perus#koulu	0.0056	haluta	0.0043	jatko	0.0031
ammatti#korkeakoulu	0.0091	oppia	0.0054	tarve	0.0042	riittävä	0.0031
tutkimus	0.0083	tulevaisuus	0.0054	oppi#laitos	0.0038		
oppilas	0.0083	tarvita	0.0053	esi#opetus	0.0037		

Kuvio 5. Esimerkkituloste yhdestä 30-aiheisen LDA-mallin tuottaman aiheen 50 merkitsevimmästä sanasta. Aiheen sanastoa esiintyy 22,57 prosentissa analysoiduista puheista ($p > 0,01$), ja merkitsevin aihe se on 12,12 prosentissa puheenvuoroista.

LDA-aiheen merkittävimpien sanojen analyysiin on hienovaraisempiakin mittareita ja ihmistarkastelijan työtä helpottavia visualisointivälineitä, kuten pyLDAvis (Sievert ja Shirley 2014), joilla voi vertailla sanojen merkitystä yhden aiheen sisällä sekä koko korpuksen tasolla.

Aiheiden nimeäminen arkikielisesti ymmärrettävällä tavoin on tärkeää, jotta LDA-malleilla saaduista tuloksista pystytään keskustelemaan. Esimerkiksi jos täysistuntokeskustelusta tuotetun aiheen viisi merkitsevintä sanaa ovat 'poliisi', 'turvallisuus', 'valvonta', 'resurssi', ja 'rikollisuus', on luontevaa olettaa, että kyseessä on puhe poliisiasioista. Näiden jälkeen voivat tulla kuitenkin suurella merkitsevyydellä myös 'potilasturvallisuus' ja 'tulli', sillä samalla sanastolla käsitellään muitakin kuin poliisiasioita. LDA voi tuottaa laskennallisesti mielekkäitä aiheita, joiden merkitys ei kuitenkaan avaudu ihmislukijalle. Tosin joskus sanalista voi antaa lukijalle jopa paremman käsityksen aiheen sisällöstä kuin abstrakti nimi (Aletras ym. 2017).

LDA tuottaa usein myös ihmislukijan näkökulmasta satunnaisia aiheita eli niin sanottuja roska-topikkeja, tai kuten Laaksonen ja Nelimarkka (2018) muotoilevat, "aiheita, jotka eivät ole semanttisesti tulkittavissa". Tällaisten tunnistaminen on tarpeen, jotta tulosten jatkoanalyysissa tai viimeistään tulosten raportoinnissa ne voidaan huomioida, esimerkiksi jättää pois aineistosta tai sivuuttaa raportoinnissa kohinana.

Mikäli laskentakapasiteettia on riittävästi, on yleensä järkevintä laskea useita LDA-malleja ja valita niistä ihmislukijoiden tarkasteluun laskennallisesti parhaat esimerkiksi C_V-koherenssien perusteella. Eri mallien yhdenmukaisuuden tarkastelun tunnusluvut ja menetelmät ovat kehittyvä tutkimusala, ja jatkossa, kun laskentakapasiteettia on riittävästi, täysistuntokeskustelumallien laskennallinen vertailu tulee ajankohtaiseksi. Jäljempänä esittelemme kunta-asioihin keskittyvän aiheen, jollainen esiintyi johdonmukaisesti kaikissa malleissa. Eräs ajankohtainen tutkimuskysymys on, miten tällaiset säännöllisesti esiintyvät aiheet pystytään löytämään laskennallisesti (ks. esim. Belford ym. 2018). Toistaiseksi eteneminen niin sanotusti tutkijoiden näppituntuman varassa on vaikuttanut mielekkäälle työskentelytavalle.

AIEMPIA LDA-MALLIEN LAATULUOKITUKSIA

Kuinka LDA-aiheiden laatua voi analysoida ja raportoida johdonmukaisesti? Mimno ym. (2011) sekä Chuang ym. (2013) ovat luoneet tilastollisten mittarien kehitystyön yhteydessä toisistaan hieman poikkeavat luokitukset, jotka toimivat silloin, kun käytössä on valmis vertailuluokitus tai aineisto on valikoidumpaa jo ennen LDA-analyysia. Luokitukset toimivat mallina omallemme.

Mimno ym. (2011, 263–264) käyttivät National Institutes of Healthin 300 000 rahoituspäätöksestä laadittua 500-aiheista mallia. Asiantuntijat valitsivat 500 aiheesta 148 sellaista, joiden sanastosta heillä oli asiantuntemusta, ja he arvioivat aiheiden laatua 30 todennäköisimmän sanan perusteella. Osuvuus luokiteltiin ensin kolmeen luokkaan: hyvä, keskitasoinen tai huono. Keskitasoisten ja huonojen osalta luokitusta tarkennettiin täsmentämällä ongelman syy yhdellä tai useammalla seuraavista luokista:

- **Chained**: sanat liittyvät toisiinsa jollakin tavalla, mutta kaikki parit eivät ole mielekkäitä.
- **Intruded**: joko a) kaksi tai useampia yhtenäisiä sanajoukkoja, jotka eivät liity toisiinsa, tai b) muuten hyvä sanajoukko, mutta sisältää muutamia asiayhteyteen kuulumattomia ”tunkeutujasanoja”
- **Random**: mitään mielekästä yhteyttä ei ilmene kuin korkeintaan muutaman sanan välillä
- **Unbalanced**: sanajoukko liittyy toisiinsa mielekkäällä tavalla, mutta koostuu erittäin yleisistä termeistä ja erikoisalasanastosta.

Intruded-luokka sisältää kaksi alaluokkaa, joista kahden tai useamman yhtenäisen sanajoukon muodostamasta aiheesta käytetään monissa LDA-analyyseissa nimitystä *mixed topic*.

Chuang ym. (2013) vertailivat LDA:n tuottamia sanastoja asiantuntijoiden laatimaan luokitukseen viidellä luokalla:

- **Resolved**: LDA:n tuottama sanasto vastaa asiantuntijoiden tuottamaa luokkaa yksi yhteen
- **Junk**: LDA:n tuottama aihe ei vastaa mitään luokkaa asiantuntijaluokituksessa
- **Fused**: LDA:n tuottama aihe vastaa kahta tai useampaa luokkaa asiantuntijaluokituksessa
- **Missing**: asiantuntijaluokituksen luokalle ei ole vastinetta LDA-aiheissa
- **Repeated**: kaksi tai useampia LDA-aiheita vastaa yhtä asiantuntijaluokituksen luokkaa.

Chuangin ym. (2013) luokitus toimii, kun käytettävissä on valmis luokitus, ja tavoitteena on tehdä sitä vastaava LDA-malli, tai kun laskennallisia malleja kehitetään valmiiksi luokitellun korpuksen avulla. Mimnon ym. (2011) luokitus soveltuu aineistolähtöiseen analyysiin. Luokilla on selviä yhtymäkohtia: *junk* ja *random* ilmenevät sanojen joukkona, jossa sanat eivät liity yhteen ihmislukijalle mielekkäällä tavalla. *Fused*, *mixed* ja *intruded* ilmenevät ihmislukijalle samankaltaisina sanajoukkoina, jotka ihminen erottelisi useampaan erilliseen joukkoon, joista yksi voi olla eräänlainen jäännöserä sekalaisille sanoille.

Mimnon ym. (2011) aineisto koostui lääketieteen rahoitushakemuksista ja sisälsi siksi paljon erityisalasanastoa, ja lisäksi heidän tutkimuskysymystensä kannalta suurin osa aiheista oli mielekästä jättää arvioimatta. Parlamenttipuheesta tuotettujen LDA-aiheiden analyysiin on kuitenkin mielekkäämpää käyttää luokitusta, joka soveltuu kaikenlaisiin aiheisiin eikä vain erityisalasanaston perusteella valikoituihin.

UUSI LUOKITUSTAPA

Käytämme luokitusta, jonka määrittelimme Chuangin ym. (2013) ja Mimnon ym. (2011) edellä esitettyjen luokitusten pohjalta. Koko luokitteluohje on liitteenä A (artikkelin julkaisusivulla). Luokituksen etuna on, ettei se edellytä aineiston perkaamista ennen LDA-analyysia, vaan se antaa keinon viestiä erilaisten aihetyyppien ominaisuuksista ja yksittäisten aiheiden mahdollisista puutteista.

Määrittelimme seitsemän luokkaa, joihin ihmislukija sijoittaa aiheet käyttämänsä analyysimetodin mukaan. Käytimme metodinamme edellä kuvailemaamme top 50 -sanastoa, mutta luokitus soveltuu myös muiden sanamäärien sekä visualisointityökalujen käyttäjille.

- **Laadukas**: aihe vastaa yhtä ihmislukijalle mielekästä teemaa.
- **Läheiset**: erilliset aiheet, jotka liittyvät toisiinsa jollain tapaa. Esimerkiksi toimeentulotuki ja lapsilisä ovat yhteiskunnan tukia, mutta kuitenkin erilaisia tukia. Luokan käyttö riippuu myös aiheiden määrästä: jos kokonaisaihemäärä on pieni, voidaan tulkita, että kaikenlainen puhe yhteiskunnan tukimuodoista sopii yhteen aiheeseen.
- **Erilliset**: aihe vastaa kahta tai useampaa ihmislukijalle mielekästä teemaa.
- **Tunkeutunut**: aihe vastaa pääpiirteissään yhtä ihmislukijalle mielekästä teemaa, mutta todennäköisimpien sanojen joukossa esiintyy useita sanoja, jotka eivät sovi asiayhteyteen.
- **Satunnainen**: aiheen sanasto ei muodosta ihmislukijalle mielekästä kokonaisuutta.
- **Ketjuuntunut**: aiheessa esiintyy kaksi erillistä teemaa, jotka liittyvät toisiinsa jonkin yhdistävän sanajoukon kautta. Esimerkiksi puheet terrorismista ja lasten kotihoidosta saattavat luokittua samaan aiheeseen, koska molemmissa puhutaan tukemisesta. Tosin tällaisia aiheita esiintyy yleensä varsin niukasti.
- **Vinoutunut:** aiheessa yksi tai muutama sana saavat erittäin suuren todennäköisyyden, ja muilta osin todennäköisyydet ovat pieniä. Esimerkiksi sanat ’kannattaa’ ja ’ehdotus’ esiintyvät todennäköisyydellä > 0,1, ja muuten sanasto liittyy vain löyhästi johonkin teemaan tai on satunnaista. Jos sanasto liittyy johdonmukaisesti samaan teemaan, kyseessä on laadukas aihe.

- **Yleissanoja**: aihe sisältää kielessä yleisesti esiintyviä sanoja, esimerkiksi 'paljon', 'mennä', 'silloin', 'puhua', 'ihminen', 'tietää' jne., eikä joukossa ole merkittävässä määrin erityisalasanastoa.

Luokituksen tarkoituksena on analysoida, 1) ovatko C_V-koherenssi ja ihmislukijoiden arviot yhteneviä ja 2) ovatko muodostetut luokat käyttökelpoisia muissa analyyseissa. Ainakin satunnaiset aiheet ovat käyttökelvottomia, ja analyysissa täytyy ratkaista, millä tavoin ne jätetään sivuun aineistosta tai huomioidaan pelkkänä kohinana.

Tässä luokituksessa LDA-mallin tavoitteena on tuottaa yhtä ihmislukijalle mielekästä ja järkeenkäyvää teemaa vastaava aihe. Siksi käytämme parhaasta luokituksestamme termiä "laadukas". Aina yhdelle teemalle ei mallinnu vain yhtä aihetta, vaan niitä saattaa olla useita, minkä huomiotta jättäminen on luokituksemme heikkous. Valinta on työmääräekonominen: jos ihmiskoodaaja tarkastelisi toistumista, pitäisi joko muistaa kaikkien mallin aiemmin tarkasteltujen aiheiden sisältö tai kerrata jokainen luokka, jos toistuminen on mahdollista, mikä veisi paljon aikaa ja saattaisi aiheuttaa inhimillisiä virheitä. Lisäksi jatkoanalyyseissa on usein mahdollista yhdistää toistuneet luokat.

Tulosten raportoimiseksi on yleensä tarpeen nimetä laskennalliset aiheet ihmislukijoille helposti ymmärrettävällä tavalla. Esimerkiksi jos aiheessa esiintyvät sanat 'varuskunta', 'armeija', 'reservi', 'kalusto', 'asevelvollisuus', on luontevaa nimetä aihe puolustuspolitiikaksi, vaikka sana 'puolustuspolitiikka' ei olisikaan merkitsevimpien sanojen joukossa. Käytäntönämme on nimetä aihe käyttäen teemaa yhdistävää abstraktia kokoavaa käsitettä, mikäli sellainen on suomen kielessä, tai ellei sopivaa käsitettä ole, kuvailemalla aihetta muutamilla aiheen merkitsevimmillä sanoilla.

Luokkien laadun vertailemiseksi järjestimme luokat viisiportaiselle asteikolle huonoimmasta parhaimpaan:

1 Satunnainen
2 Vinoutunut
3 Erilliset, ketjuuntunut tai tunkeutunut
4 Läheiset
5 Laadukas

Yleissanoja-luokka ei sovellu käytettäväksi vastaavalla paremmuusasteikolla. Useimmiten kyseinen luokka sisältää johdonmukaisesti pelkästään hyvin tavallisia sanoja ja on sikäli laadukkaasti mallintunut, mutta toisinaan joukossa on hieman harvinaisempia sanoja, minkä vuoksi luokan C_V-koherenssiarvoissa on huomattavia eroja. Jos aiheeseen on sekoittunut sekä yleissanastoa että erityisalasanastoa, oikea luokitus on satunnainen.

Järjestyksen ääripäät ovat selkeät: laadukas tai satunnaisuuden vuoksi täysin kelvoton toimivat selkeinä vertailukohtina muille. Laatuluokat 2–4 ovat järjestykseltään tulkinnanvaraisia. Esimerkiksi luokka vinoutunut voi joissain tapauksissa olla aivan käyttökelpoinen, jos halutaan analysoida jonkin tietyn sanan yhteydessä esiintyvää puhetta. Ihmislukijan arvioissakin tarvitaan aina rajankäyntiä, kun pitää miettiä, kuinka monta tunkeutujasanaa tekee yksiteemaisesta aiheesta tunkeutuneen tai ovatko kaksi politiikan teemaa niin läheisiä toisilleen, että

niitä voi pitää laadukkaana luokkana. Esimerkiksi arkikielessä puhutaan usein ”ulko- ja turvallisuuspolitiikasta” yhtenä kokonaisuutena, mutta LDA-malleissa näille näyttäytyy monta sanastoa liittyen esimerkiksi puolustusvoimiin varuskuntien ja maamiinojen yhteydessä, kansainväliseen kriisinhallintaan, Natoon ja YK:n ja EU:n turvallisuuspolitiikkaan sekä näiden esiintymisiin yhdessä LDA-aiheessa. Riippuu kuitenkin tutkimuskysymyksestä, tarvitaanko laadukkaista aiheista koostuva malli vai riittääkö joidenkin tiettyjen teemojen laadukas mallintuminen.

EDUSKUNTA-LDA-MALLIEN IHMISLUKIJA-ANALYYSI

Laskettuja LDA-malleja on tässä vaiheessa arvioinut toinen kirjoittajista ja kolme muuta koodaajaa[10], jotka tutustuivat LDA-malleihin luokitteluohjeen, yleisluontoisen artikkelin (Blei 2012) ja lyhyen perehdytyksen avulla.

Tulkinnoissa on huomattavia eroja. Luokittelijat käyttivät luokkia monin paikoin eri tavoin (taulukot 1 ja 2). Satunnaisten ja yleissanojen osuudet ovat melko samansuuruisia. Sen sijaan tutkija tulkitsi luokat laadukkaiksi paljon harvemmin ja käytti huomattavasti enemmän luokkia 2–4, joissa LDA-aiheesta pystytään saamaan informaatiota, vaikka sen tulkintaan liittyy hankaluuksia.

Laatu-luokka	**Luokitus**	**Luokituksia tutkija**		**Luokituksia koodaaja 1**		**Luokituksia koodaaja 2**	
1	Satunnainen	139	14,2 %	132	13,5 %	130	13,3 %
2	Vinoutunut	210	21,4 %	51	5,2%	52	5,3 %
3	Erilliset/tunkeut./ketj	200	20,4 %	95	9,7 %	170	17,3 %
4	Läheiset	131	13,4 %	74	7,6 %	112	11,4 %
5	Laadukas	236	24,1 %	566	57,8 %	433	44,2 %
	Yleissanoja	64	6,5 %	62	6,3 %	83	8,5 %
	Yhteensä	**980**	**100 %**	**980**	**100 %**	**980**	**100 %**

Taulukko 1. Käytettyjen luokkien määrät, kaikki mallit yhteensä.

Vinoutuneet luokat esiintyivät pääosin 200 aiheen malleissa (vrt. taulukko 2). Tämä on odotettavaa, koska aihemäärän lisääntyessä kielelliset ilmiöt pääsevät herkemmin esiin. Esimerkiksi sanapari ’kannattaa’ ja ’ehdotus’ muodostavat yleisyytensä vuoksi vinoutuneen luokan helposti jo pienehköllä aihemäärällä, mutta ’kantaa’ ja ’vastuu’ pääsevät esiin vasta suuremmalla aihemäärällä. Suurin osa vinoutunut-luokan käytöstä perustui kuitenkin yhden sanan saamaan suureen todennäköisyyteen.

Laatu-luokka	Luokitus	Luokituksia tutkija		Luokituksia koodaaja 1		Luokituksia koodaaja 2		Luokituksia koodaaja 3	
1	Satunnainen	67	11,6 %	58	10,0 %	79	13,6 %	57	9,8 %
2	Vinoutunut	57	9,8 %	2	0,3 %	26	4,5 %	46	7,9 %
3	Erilliset/tunkeut./ketj.	130	22,4 %	73	12,6 %	96	16,6 %	109	18,8 %
4	Läheiset	99	17,1 &	54	9,3 %	71	12,2 %	67	11,6 %
5	Laadukas	190	32,8 %	365	62,9 %	261	45,0 %	255	44,0 %
	Yleissanoja	37	6,4 %	28	4,8 %	47	8,1 %	46	7,9 %
	Yhteensä	**580**	**100 %**	**580**	**100 %**	**580**	**100 %**	**580**	**100 %**

Taulukko 2. Käytettyjen luokkien määrä, kun 200-aiheisia malleja ei huomioida (T30, T50, T60, T70, T80; a ja b).

Krippendorffin alfa-kertoimet[11] jäivät välille 0,39–0,62 mallien keskiarvon ollessa 0,52 ja keskihajonnan 0,076, mikä ei ole yhdenmukaisuudeltaan tyydyttävällä tasolla kuin muutamien mallien kohdalla. Mallien välillä esiintyy huomattavan suurta vaihtelua, vaikka käytetty ohjeistus ja mallinnustapa ovat pysyneet samana. Eräs eroja aiheuttava tekijä on se, että yksi neljästä luokittelijasta arvioi aiheet laadukkaiksi muita useammin. Erityisesti jos joku arvioi aiheen vinoutuneeksi ja toiset laadukkaaksi, päädytään järjestysasteikon eri päätyihin.

Toteutettu luokittelutyö tehtiin kirjallisten ohjeiden perusteella, ja koodaajien alkuperehdytys oli suppea. Luultavasti perusteellisempi opastus ja hankalasti luokiteltavien rajatapausesimerkkien käsittely yhtenäistäisivät luokituksia. Toisaalta ihmiset tulkitsevat LDA-aiheita usein eri tavoin, ja aiempi tuntemus teemasta voi helpottaa näkemään yhteyksiä sanajoukoissa, joissa se muille lukijoille ei olisi ilmeistä. Osa koodaajista kiinnitti luokitellessaan huomiota ajankohtaisaiheisiin etsimällä yhteistä nimittäjää viimeisimpien vuosien päivänpolttavista teemoista, toiset puolestaan keskittyivät puhtaammin sanojen semanttisiin yhteyksiin jättäen avoimeksi mahdollisuuden, että samaa aihetta on saatettu käsitellä useissa yhteyksissä. Esimerkiksi samalle aiheelle annettiin nimeksi ”tasa-arvoinen avioliitto” ja ”lakivaliokunta, sukupuoli, avioliitto, adoptio”.

Käytännön sovelluksena LDA-mallin laadukkuuden arvioija voi vastata seuraaviin kysymyksiin (Nikolenko ym. 2017): A) Näetkö perusteen, miksi aiheen sanat esiintyvät yhdessä; onko jokin ilmeinen semanttinen syy, joka yhdistää aiheen sanoja? B) Tunnistatko asioita tai tapahtumia (yksittäisiä tai toistuvia), joihin aihe liittyy?

IHMISLUKIJALUOKITUKSEN JA C_V-KOHERENSSIN YHTEYS

Laskennallisen LDA-aiheen laadukkuuden ja ihmislukijoiden arvioiden välillä vallitsee yhteys, mutta se ei ole aivan suoraviivainen. Laskimme C_V-koherenssit 10, 20, 30 ja 50 merkitsevimmän sanan perusteella, näiden keskinäiset korrelaatiot sekä korrelaation ihmisluokittelijoiden luokitusten keskiarvolle (taulukko 3). Yleissanoja-luokka kirjattiin puuttuvaksi havainnoksi, vaikka pääsääntöisesti siihen oli mallintunut yksinomaan yleissanastoa ja LDA toiminut sikäli toivotulla tavalla.

	Laatuluokka	**C_V10**	**C_V20**	**C_V30**
C_V10	0,604			
C_V20	0,621	0,915		
C_V30	0,620	0,862	0,964	
C_V50	0,597	0,786	0,901	0,960

Taulukko 3. Ihmislukijoiden TOP 50 -laatuarvioiden ja CV-koherenssien Spearman-korrelaatiot. Kaikki mallit, 30–80-aiheisista malleista 4 luokittelijaa, 200-aiheisesta 3. Korrelaatioiden merkitsevyys > 0,001; n = 3920, paitsi laatuluokka N = 3267, koska yleissanoja-luokkaa ei huomioitu.

Ihmislukijoiden tulkinnat 50 merkitsevimmän sanan ja niiden todennäköisyyksien perusteella korreloivat melko voimakkaasti C_V-koherenssiarvojen kanssa. Korrelaatioissa oli kuitenkin huomattavia eroja ihmisluokittelijoiden välillä (taulukko 4). LDA-malleihin ja eduskuntapuheeseen perehtyneen kirjoittajan arviot korreloivat huomattavasti voimakkaammin C_V-koherenssiarvojen kanssa kuin koodaajalla 1 ja 2, mutta voimakkain korrelaatio havaittiin koodaajalla 3. Jos 200-aiheiset mallit jätettäisiin huomiotta, korrelaatio olisi hieman heikompi koodaajilla 1 ja 2, ja kirjoittajalla (tutkijalla) likimain entisellään.

		C_V10	C_V20	C_V30	C_V50
Laatuluokka tutkija	Spearman-korrelaatio	0,676	0,687	0,686	0,656
	N	916	916	916	916
Laatuluokka koodaaja 1	Spearman-korrelaatio	0,567	0,576	0,574	0,544
	N	918	918	918	918
Laatuluokka Koodaaja 2	Spearman-korrelaatio	0,597	0,62	0,612	0,594
	N	897	897	897	897
Laatuluokka koodaaja 3	Spearman-korrelaatio	0,652	0,697	0,718	0,708
	N	536	536	536	536

Taulukko 4. CV-koherenssien ja ihmislukijaluokittelijoiden korrelaatiot, tutkija ja kolme koodaajaa. Yleissanoja-luokkaa ei huomioitu.
Korrelaatioiden merkitsevyys > 0,001. Koodaaja 3 luokitteli vain 30–80-aiheiset mallit.

Analyysin perusteella C_V-koherenssia voi pitää käyttökelpoisena eduskunta-LDA-mallien analyysivälineenä. Nyt noin 20–30 ensimmäistä sanaa huomioiva koherenssiluku näyttää korreloivan voimakkaimmin ihmislukijoiden arvioiden kanssa. Mitä suurempi määrä aiheita mallissa on, sitä paremmin toimivat pienempiä sanamääriä huomioivat C_V-koherenssiluvut.

Taulukosta 5 havaitaan, että suuremmilla aihemäärillä sekä ihmiskoodaajien luokitukset että C_V-koherenssit ovat keskimäärin pienempiä, kun tarkastellaan koko mallia. Suurilla aihemäärillä malleissa on kuitenkin myös lukuisia erittäin laadukkaita aiheita, ja tutkijan täytyy tehdä valinta suuren aihemäärän tuoman erottelukyvyn ja pienemmän aihemäärän helpomman yleistettävyyden ja raportoitavuuden välillä.

Malli	ihmisluokittelija-ka.	C_V10	C_V20	C_V30	C_V50
T30a	4,17	0,724	0,734	0,743	0,744
T30b	4,24	0,746	0,762	0,774	0,788
T50a	4,04	0,744	0,731	0,736	0,737
T50b	4,18	0,746	0,746	0,753	0,751
T60a	3,81	0,724	0,715	0,709	0,714
T60b	3,83	0,727	0,707	0,698	0,690
T70a	3,84	0,726	0,696	0,684	0,672
T70b	3,80	0,717	0,692	0,682	0,675
T80a	3,67	0,716	0,694	0,681	0,666
T80b	3,69	0,715	0,681	0,667	0,652
T200a	3,23	0,666	0,619	0,582	0,563
T200b	3,14	0,660	0,616	0,583	0,565

Taulukko 5. LDA-mallien vertailu ihmisluokittelija-arvioiden ja CV-koherenssien keskiarvoilla.

Koherenssiluvuissa samoilla aihemäärillä ym. parametrien arvoilla lasketuissa malleissa on yleensä varsin pieniä eroja. Taulukossa 5 suurimmat erot esiintyvät 30- ja 50-aiheisissa malleissa, ja ne ovat samansuuntaisia ihmisluokittelijoiden kanssa. Kuten aiemmin todettiin, kun laskentakapasiteettia on riittämiin, LDA-malleja kannattanee laskea lukuisia ja ottaa jatkotarkasteluun ne, joiden C_V-koherenssiarvot ovat suurimmat. LDA-aiheiden koherensseissa on suuria eroja mallien sisällä, ja keskimäärin heikompia arvoja saanut malli voi sisältää monia käyttökelpoisia aiheita.

KUNTAPUHE: ESIMERKKI LAADUKKAASTI MALLINTUVASTA AIHEESTA

Taulukossa 6 on listattuna kunta-aiheiden 10 merkitsevintä sanaa merkitsevyys-järjestyksessä. Tilan säästämiseksi sanoja on raportoitu vain 10 kappaletta; 50 merkitsevintä sanaa todennäköisyyksineen ovat liitteenä B (artikkelin julkaisusivulla).

Koodaajat tulkitsivat aiheet kuntateemaisiksi, ja ne olivat lähes kaikkien luokittelijoiden mukaan laadukkaasti mallintuneita miltei kaikissa malleissa (84 prosenttia luokituksista). Joissain tapauksissa luokittelijat tulkitsivat aiheen luokkaan läheiset, esimerkiksi aiheessa T70a#34 yksi luokittelija piti teemoja ”kuntatalous, kunnan tehtävät” läheisinä, kun taas muut kolme luokittelijaa tulkitsivat puheen aiheen laadukkaaksi. He nimesivät aiheen samansuuntaisesti, mutta eri asioita painottaen: ”kunta, tehtävä, valtionosuus”, ”kunnat peruspalvelujen järjestäjänä” ja ”valtionosuudet”.

Kunta-aihe toistui selkeästi erottuvana jokaisessa mallissa. Samankaltaisia asioita käsiteltiin muissakin aiheissa, esimerkiksi sote-asiat erottuivat omina aiheinaan, ja toisaalta kuntasanastoa esiintyi muissakin aiheissa.

T30a #27	**T30b #3**	**T50a #41**	**T50b #30**	**T60a #12**	**T60b #14***
kunta	kunta	kunta	kunta	kunta	kunta
palvelu	palvelu	palvelu	palvelu	palvelu	valtion#osuus
valtion#osuus	terveyden#huolto	järjestää	valtio	valtio	kunta#talous
valtio	sosiaali	valtion#osuus	valtion#osuus	valtion#osuus	valtio
järjestää	järjestää	uudistus	järjestää	järjestää	palvelu
kunta#talous	valtion#osuus	kuntalainen	uudistus	uudistus	hallitus
tehtävä	julkinen	kunnallinen	kunta#talous	kunta#talous	kuntalainen
suuri	valtio	sosiaali	suuri	suuri	perus#palvelu
perus#palvelu	kunta#talous	perus#palvelu	alue	perus#palvelu	tehtävä
kunnallinen	suuri	terveys#palvelu	tehtävä	tehtävä	suuri
kuntalainen	turvata	tehtävä	kunnallinen	kunnallinen	asukas
julkinen	kunnallinen	turvata	yhteis#työ	turvata	kunta#uudistus
turvata	tarvita	suuri	perus#palvelu	kuntalainen	kunta#liitos

T70a #34*	**T70b #11**	**T80a #63****	**T80a #64***	**T80b #24**	**T200a #123****
kunta	kunta	kunta	malli	kunta	kunta
palvelu	palvelu	palvelu	maa#kunta	palvelu	valtio
valtion#osuus	valtion#osuus	valtion#osuus	paikka#kunta	valtio	valtion#osuus
valtio	valtio	valtio	keskittää	valtion#osuus	kunta#talous
turvata	kunta#talous	kunta#talous	kunta#uudistus	kunta#talous	perus#palvelu
järjestää	järjestää	perus#palvelu	vahva	järjestää	tehtävä
kunta#talous	kunnallinen	järjestää	kunta#liitos	kunnallinen	rahoitus
perus#palvelu	kuntalainen	suuri	jyväskylä	perus#palvelu	valtion#osuus#järjestelmä
julkinen	asukas	asukas	kunta#rakenne	kuntalainen	velvoite
sosiaali	suuri	sosiaali	perus#kunta	suuri	talous
uudistus	tehtävä	kuntalainen	keuruu	asukas	kela#maksu
tehtävä	perus#palvelu	turvata	lähi#palvelu	hallitus	meno
suuri	hallitus	talous	alue	tehtävä	kustannus

T200a #188	**T200b #45**	**T200b #54****
kunta	kunta	palvelu
palvelu	valtion#osuus	kunta
uudistus	kunnallinen	kunta#talous
asukas	kuntalainen	turvata
kunnallinen	valtio	perus#palvelu
järjestää	tehtävä	suuri
kuntalainen	järjestää	kansa#lainen
kunta#uudistus	asukas	tuottaa
kunta#liitos	kunta#liitos	sairaan#hoito#piiri
vahva	pieni	järjestää
valtuusto	suuri	tarvita
terveys#palvelu	palvelu	saatavuus
suuri	itsehallinto	maa

* *Yksi luokittelija käytti muuta luokkaa kuin laadukas.*

** *Kaksi luokittelijaa käytti muuta luokkaa kuin laadukas.*

Taulukko 6. LDA-mallien kunta-aiheiden kymmenen merkitsevintä sanaa. Kullekin aihemäärälle (merkintä T[nro]) on laskettu kaksi mallia, a ja b. Malleissa T80a, T200a ja T200b kunta-aiheita esiintyi kaksi kunta-aihetta.

Taulukossa 7 on niiden aiheiden määrä, joissa kunta-aihetta esiintyy, sekä ne aiheet, joissa kunta-aihe on merkitsevämpi kuin muut aiheet. Jaottelu on sikäli karkea, että puheenvuorot muodostuvat useista LDA-aiheista, eikä nyt erotella sitä, onko kunta-aihe läsnä pienellä todennäköisyydellä vai kenties lähes yhtä suurella kuin merkitsevin aihe. Samaten siitä, että aihe on merkitsevin, ei voida päätellä, onko koko puheenvuoro puhtaasti kuntapuhetta, sillä yhdessä puheenvuorossa voi olla lukuisia aiheita, jotka ovat melkein yhtä merkitseviä. Nyansoidumpien esitystapojen kehittely onkin yksi analyysin haasteista.

Malli (aihe #nro)	**Puheenvuoroja, joissa aihe esiintyy siten, että p > 0,01**		**Puheenvuoroja, joissa aihe on merkitsevin**	
T30a #27	57045	31.46 %	3235	5.67 %
T30b #3	58835	32.45 %	5462	9.28 %
T50a #41	32104	17.70 %	1596	4.97 %
T50b #30	52767	29.10 %	5522	10.46 %
T60a #12	50496	27.85 %	5245	10.39 %
T60b #14	31803	17.54 %	1765	5.55 %
T70a #34	47827	26.38 %	3167	6.62 %
T70b #11	32915	18.15 %	1775	5.39 %
T80a #63	33693	18.58 %	2537	7.53 %
T80a #64	19593	10.81 %	57	0.29 %
T80b #24	34508	19.03 %	1950	5.65 %
T200a #123	21376	11.79 %	610	2.85 %
T200a #188	15528	8.56 %	1812	11.67 %
T200b #45	26787	14.77 %	1709	6.38 %
T200b #54	22844	12.60 %	3016	13.20 %

Taulukko 7. Kunta-aiheiden esiintyminen aineistossa, puheenvuorojen määrät ja prosenttiosuudet. Ne mallit, joissa kunta-aihe toistuu, on lihavoitu.

Kunta-aiheen esiintymismäärät vaihtelevat melko suuresti eri mallien välillä ja myös malleissa, joiden aihemäärä on sama. Sanaston esiintyminen mallin muissa aiheissa vähentää ”puhtaan” kuntapuheen todennäköisyyttä. Kun aihemäärä on suuri, aihe voi toistua, ja keskeisimmät sanat, kuten ’kunta’, saattavat mallintua omaan aiheeseensa. Esimerkiksi malleissa T30a ja T30b kunta-aiheita on yhteensä likimain yhtä paljon, mutta b-mallissa kunta-aihe on useammin kaikkien merkitsevin. Tämä selittyy sillä, että kummassakin mallissa on aihe, johon kuuluu aluepolitiikka, maakunnat ja pääkaupunkiseutu, mutta mallissa T30a samaan yhteyteen on

mallintunut kuntarakenne ja siihen liittyvä sanasto. Mallin T200a aiheessa #123 sanan ’kunta’ todennäköisyys puolestaan on 0,4864, ja kyseisen aiheen muutamien merkitsevimpien sanojen jälkeen yleissanojen osuus on suuri, ja muuta kuntasanastoa on mallintunut aiheeseen #188. Teeman mahdollinen toistuminen tekee myös laadukkaasti mallintuneiden aiheiden tulkinnasta haasteellista, ja silloin kannattanee harkita aihemäärän vähentämistä.

KORRELAATIOANALYYSI: MILLOIN KAKSI AIHETTA ESIINTYY SAMOISSA PUHEENVUOROISSA?

Esiintyvätkö jotkin aiheet usein toistensa yhteydessä tai ovatko jotkin aiheet sellaisia, ettei niiden yhteydessä puhuta jostakin tietystä muusta aiheesta? LDA:n tuottama dokumenttikohtainen informaatio, kunkin dokumentin aiheiden todennäköisyydet, on aina summaltaan vakio. Siksi tavanomaiset korrelaatiokertoimet antaisivat negatiivisesti harhaisia tuloksia ja ne mahdollisesti muodostaisivat näennäisiä yhteyksiä (ks. esim. Pawlowsky-Glahn ym. 2015, 4–7). Kyse on suhteellisista osuuksista (*compositional data*), jolle ei ole vakiintunutta vastinetta suomeksi ja jota nimitämme jatkossa osuusdataksi. Jos korrelaatioanalyysi on tutkimuksen tärkein päämäärä, kannattanee kuitenkin analysoida aineisto correlated topic model -aihemallilla (Blei ja Lafferty 2007).

Vakiosumman ongelmaa on alkujaan käsitelty matemaattisen geokemian analyyseissa, mutta matemaattinen ongelma on sama kaikille tilanteille, joissa aineisto summautuu vakioksi. Esimerkiksi puolueiden paikkaosuuksien keskinäisen korrelaation tarkastelussa kohdattaisiin samanlainen haaste.

Kynčlová ym. (2017) kehittivät korrelaatiokertoimen suhteellisille osuuksille. Kertoimelle ei ole vakiintunutta suomenkielistä nimeä – Kynčlová-Hron-Filzmoser-osuuskorrelaatio olisi täsmällinen nimitys, mutta jatkossa puhumme lyhyemmin osuuskorrelaatiosta. Kerroin on laskettavissa robCompositions R-pakettiin sisältyvällä corCoDa-funktiolla. Muitakin osuuskorrelaation laskentatapoja on olemassa. Käyttämämme perustuu symmetrisiin tasapainoihin (symmetric balances), joissa korrelaatiokerroin on toivottavalla tavalla sama riippumatta siitä, missä järjestyksessä muuttujat sijoitetaan laskukaavaan. Osuuskorrelaatio eroaa tutuista Pearson-, Kendall- ja Spearman-korrelaatioista sikäli, että niissä huomioidaan koko osuusdatan sisältämä informaatio, ei ainoastaan niiden kahden muuttujan, joiden välille osuuskorrelaatiokerroin lasketaan.

LDA-mallien yhteydessä aiheiden välillä ei esiintyne erityisen suuria korrelaatioita, sillä mallinnusprosessissa johdonmukaisesti yhdessä esiintyvästä sanastosta pitäisi muodostua oma yksi aiheensa. LDA-mallissa aiheet oletetaan toisistaan riippumattomiksi. Voimakas korrelaatio olisikin merkki teeman toistumisesta, mikä ei ole mallin laadun kannalta toivottavaa. Korrelaatiokertoimet ovat kiinnostavia silloin, kun tarkastellaan osa-aineistoja, esimerkiksi esiintyykö aiheen A yhteydessä aihe B erilaisessa määrin eri ryhmillä.

Jäljempänä vertailemme kiinnostuksemme kohteena olevien aiheiden korrelaatioita valtiopäivävuosittain. Jakamalla aineisto aikajaksoihin LDA-mallin laskemisen jälkeen pystytään LDA:ta käyttämään samaan tapaan kuin dynaamisia aihemalleja, joiden laskenta perustuu dokumenttien esittämisjärjestykseen (vrt. Quinn ym. 2010, Greene ja Cross 2017). Aikajaksottaisessa

tarkastelussa on kuitenkin tärkeää huomioida, että mallien oletukset ovat erilaisia. LDA lasketaan koko aineistolle, kun taas dynaamisissa malleissa oletetaan, että sanasto muuttuu ajan kuluessa.

KUNTAPUHEEN YHTEYS SOTE-, DEMOKRATIA- JA POLITIIKKA-AIHEISIIN

Kuntapuhe korreloi LDA-malleissa selvästi muutamien aiheiden kanssa. Osuuskorrelaatiokertoimien arvot jäävät varsin pieniksi, mutta valtiopäivävuosittain ne eroavat selvästi. Kertoimien pienuus on odotettavaa, sillä LDA-aiheiden tulisikin olla toisistaan selvästi erottuvia, ja kaikissa puheissa yhtenäisesti esiintyvä sanasto mallintuisi samaan aiheeseen. Kiintoisinta on tarkastella, kuinka korrelaatiot eroavat aineiston osajoukkojen välillä, kuten eduskuntaryhmittäin, valtiopäivävuosittain, sukupuolittain tai hallitus–oppositio-aseman mukaan.

Seuraavat korrelaatiot on laskettu R:n robCompositions-paketin corCoDa-funktiolla (Kynčlová ym. 2017). LDA:n tuottamien mallien aiheissa todennäköisyyksien jakaumat ovat erittäin vinoja, ja niiden mediaani on yleensä lähellä nollaa, vaikka keskiarvot olisivat nollaa suurempia.

Esimerkkinä kuntapuheen kanssa korreloivista aiheista mallissa T50a kuntapuhe (aihe #41) korreloi positiivisimmin seuraavien aiheiden kohdalla:

#45 pk-seutu, asuntopolitiikka, kuntatalous (keskiarvo 0,203, keskihajonta 0,048)
#38 sote (ka 19,4, kh 0,043)
#26 vanhukset, omaishoito, Kela (ka 0,175, kh 0,036)
#11 alueellinen kehitys (ka 0,135, kh 0,049)
#13 hallitusohjelma, talouskasvu, leikkaukset (ka 0,135, kh 0,042)
#37 demokratia, vaalit, politiikka (ka 0,102, kh 0,096).

Korrelaatiot ovat suurimmillaankin 0,2:n ja 0,3:n välillä, maksimina 0,33. Korrelaatioissa on nähtävissä valtiopäivävuosittaista vaihtelua samojen aiheiden välillä. Osin korrelaatio selittynee muutamien kuntasanojen päätymisellä muihin kuin itse kunta-aiheeseen: ’kuntatalous’ on myös aiheessa #45 ja ’kuntalaki’ aiheessa #37. Kaiken kaikkiaan malli T50a sopii hyvin kunta- ja sote-asioiden analysointiin, koska edellä mainittuja kahta keskeistä sanaa lukuun ottamatta sekä kunta- että sote-sanasto ovat mallintuneet siististi omiksi aiheikseen.

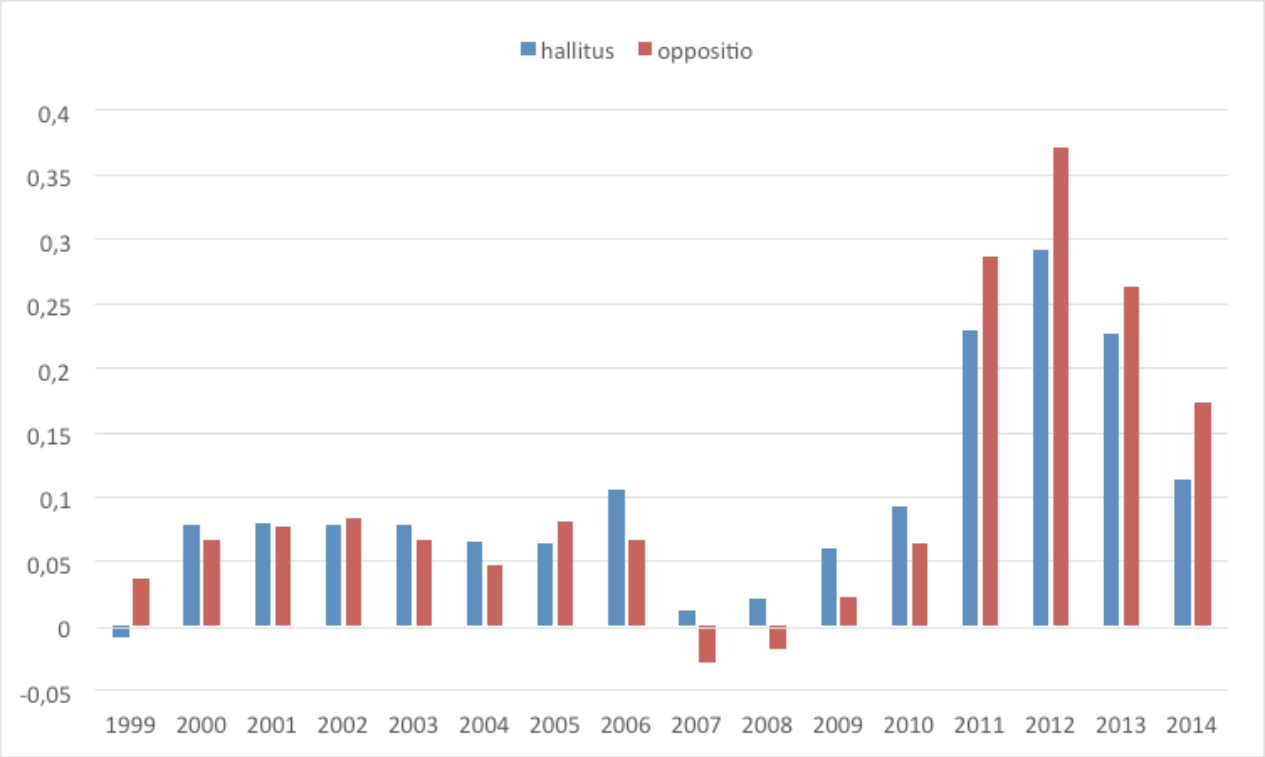
Aiheiden #37 ”demokratia, vaalit, politiikka” ja kunta-aiheen #41 valtiopäivävuosittaisten korrelaatioiden keskihajonta on huomattavasti suurempi kuin muilla aiheilla. Näiden aiheiden korrelaatio oli suurimman aikaa varsin pientä, mutta valtiopäivävuosina 2011–2013 paljon voimakkaampaa. Valtiopäivävuodelle 2012 osuu T50a-mallin voimakkain valtiopäivävuosittainen korrelaatio 0,331. Sen sijaan sote-aihe #38 korreloi jatkuvasti positiivisesti kuntapuheen kanssa, vaikka korrelaation voimakkuus hieman vaihteleekin valtiopäivävuosien välillä. Demokratia-aiheet tulivat kiinteäksi osaksi kuntapuhetta siis vasta vuoden 2011 eduskuntavaalien jälkeen, vaikka toki ne ovat jossain määrin olleet esillä muulloinkin.

Selkeästi negatiivinen korrelaatio valtiopäivävuosittain jaoteltuna on havaittavissa aiheissa:

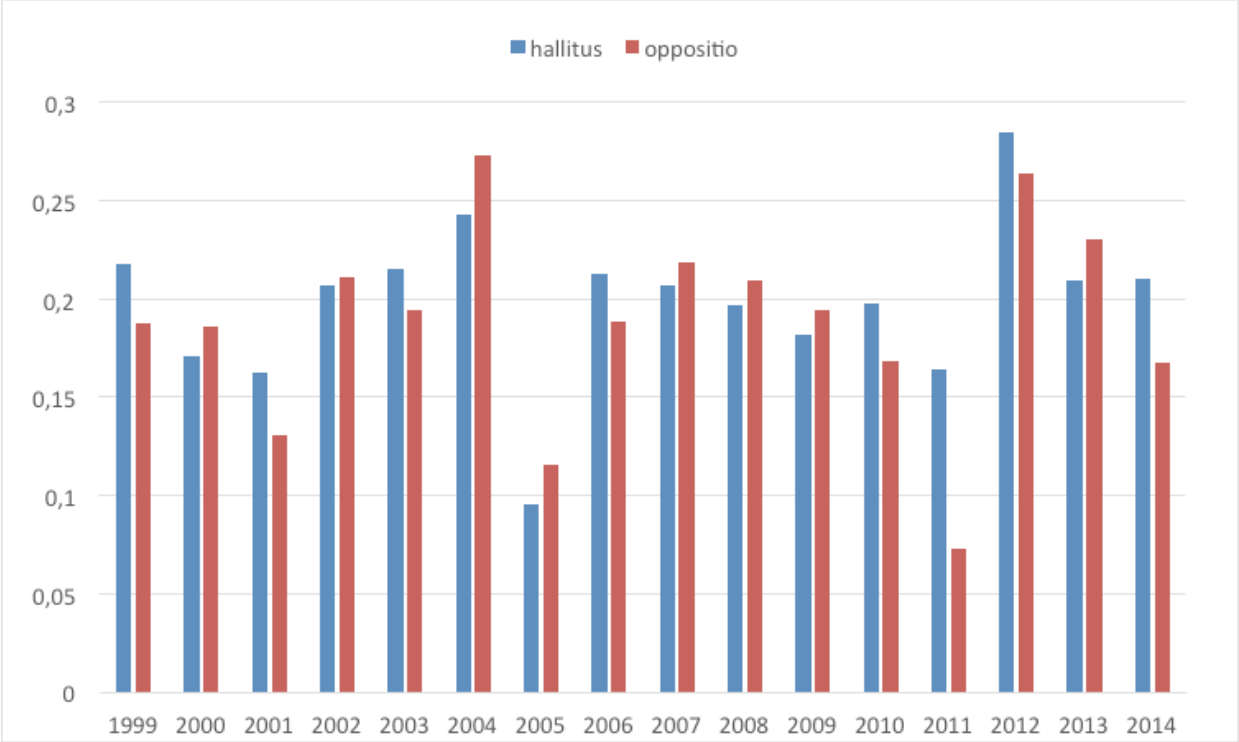
#28 ”EU” (ka -0,137, kh 0,024),
#15 ”turvallisuuspolitiikka, sopimus, Nato, Venäjä” (ka -0,103, kh 0,023),
#4 ”pohjoismaat, neuvosto, auto, ihmisoikeudet” (ka -0,101, kh 0,026),
#3 ”kauppa, harmaa talous, huumeet” (ka -0,094, kh 0,018),
#27 ”maa- ja metsätalous” (ka -0,094, kh 0,035),
#40 ”YK, kriisinhallinta, kehitysyhteistyö” (ka -0,081, kh 0,031) ja
#31 ”energiapolitiikka” (ka -0,079, kh 0,024).

Tarkasteluajanjaksolla kunta-aiheen yhteydessä ei yleensä esiintynyt perustuslakiaihetta #48 ”perustuslakivaliokunnan lausunto” (ka -0,060, kh 0,027) eikä aihetta #35 ”ylimmät valtioelimet, ulko- ja turvallisuuspolitiikka” (ka -0,064, kh 0,025), jossa perustuslakiasioita käsitellään hallitusmuodon osalta. Myös EU-aiheessa korrelaatio on negatiivinen. Kun vastaisuudessa uusimmatkin pöytäkirjat saadaan mallinnettua, on kiintoisaa nähdä, tapahtuuko asiassa muutosta.

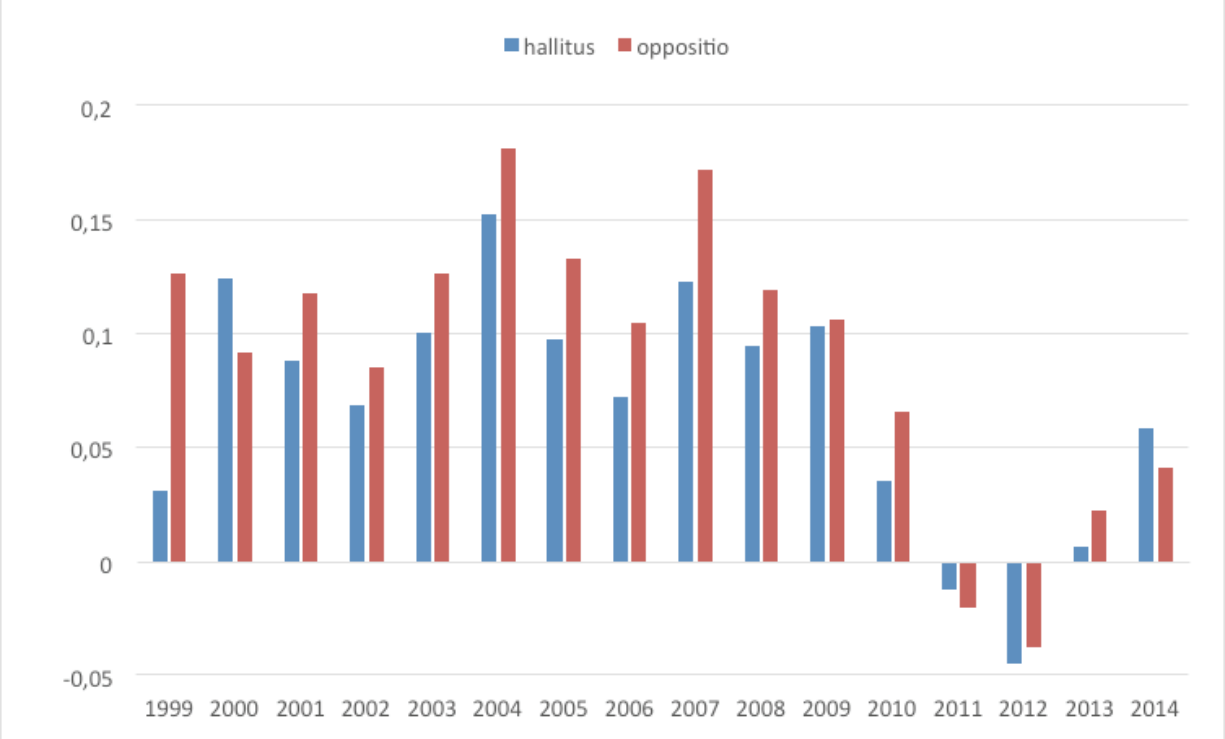
Täysistuntokeskustelussa tapahtui merkittävä muutos kunta-asioissa vuoden 2011 jälkeen. Kuvioista 6–8 on nähtävissä, että ennen vuoden 2011 vaaleja kunta-aihe korreloi budjetti-aiheen kanssa, mutta tämän jälkeen tilanne muuttui. Kuntapuheen yhteydessä alettiin puhua huomattavasti enemmän demokratiasta. Budjettipuhe jäi syrjään, ja sen korrelaatio katosi tai muuttui jopa negatiiviseksi. Sote-aiheen (T50a#38) ja kuntapuheen korrelaatio puolestaan pysyi melko vakaana koko ajanjakson, vaikka 2012 valtiopäivillä onkin havaittavissa korrelaation vahvistuminen edeltäviin valtiopäiviin verrattuna.



Kuvio 6. Kunta-aiheen (#41) ja demokratia-aiheen (#37) korrelaatio mallissa T50a valtiopäivävuosittain ja eduskuntaryhmittäin



Kuvio 7. Kunta-aiheen (#41) ja sote-aiheen (#38) korrelaatio mallissa T50a valtiopäivävuosittain ja eduskuntaryhmittäin.



Kuvio 8. Kunta-aiheen (#41) ja budjetti-aiheen (#47) korrelaatio mallissa T50a valtiopäivävuosittain ja eduskuntaryhmittäin.

Vuoden 2011 valtiopäivillä erityisesti oppositioon siirtyneen keskustan puheenvuoroissa kunta- ja demokratia-aiheet korreloivat voimakkaasti, ja vastaava muutos voitiin havaita muilla eduskuntaryhmillä seuraavana valtiopäivävuonna. Näin muodostuu vaikutelma, että keskusta olisi tuonut demokratiapuheen kuntateemaiseen täysistuntokeskusteluun ja muut ryhmät seuranneet myöhemmin hallitusryhmien pyrkiessä runnomaan läpi sittemmin hylättyä suurta kuntauudistusta. Aiemmin kunta–valtio-suhteella viitattiin rahoitusosuuksiin, kun eduskunta sääti kuntien velvollisuuksista ja valtion rahoituksesta kunnille.

LDA-aiheiden avulla perinteinen ihmislukijan analyysi on kohdennettavissa puheenvuoroihin, joissa kiinnostavat aiheet esiintyvät. Muutoksen laajempi analyysi yhdistettynä puheiden ihmislukuun ja aiheomistajuus- tai kehysteoriaan ansaitsee oman julkaisunsa, ja tässä yhteydessä kuntapuheen valtiopäivävuosittainen muutos jää esimerkiksi metodiikan mahdollisuuksista.

JOHTOPÄÄTÖKSIÄ

Edellä mallinsimme LDA-aihemalleilla eduskunnan täysistuntokeskustelut valtiopäivävuosilta 1999 (osin) – 2014. Kehitimme LDA-mallien tuottamien aiheiden ihmislukijavertailuun luokituksen, jonka toimivuus edellyttää kuitenkin luokittelijoiden kohtalaisen runsasta perehdyttämistä, sillä eri luokittelijat kiinnittävät sanaston tarkastelussa huomionsa eri asioihin, ja mallien tulkinta on pääsääntöisesti samansuuntaista, muttei täysin yhtenevää. LDA-aihemallin yhtälö ei ole laskennallisesti yksiselitteisesti ratkaistavissa, ja siksi LDA tuottaa aina hieman erilaisia malleja. Ihmislukijan analyysikehikon ja laskennallisten tunnuslukujen avulla on mahdollista arvioida, mitkä mallit on mielekkäintä ottaa jatkotarkasteluun.

LDA on moniaihemalli, mikä mahdollistaa aiheiden suhteellisten osuuksien esiintymisen tarkastelun yksittäisissä puheissa. Osuuskorrelaatioanalyysin avulla voidaan metadataa hyödyntäen tarkastella, mitkä aiheet esiintyvät samoissa puheenvuoroissa esimerkiksi eri aikoina tai eri puhujaryhmillä. Osuuskorrelaatio avaa mahdollisuuden käyttää LDA:ta dynaamisen eli aikajärjestykseen perustuvan mallin tapaan, kun samalla saavutetaan moniaihemallin edut (vrt. Greene ja Cross 2017) ja laskenta pysyy kevyenä correlated topic model -mallinnukseen verrattuna. Edellä esitetystä kuntapuhe-esimerkistä nähdään, että korrelaatiot vaihtelevat valtiopäivävuosittain ja eduskuntaryhmittäin. Kuntapuhe muutti luonnettaan raha- ja valtionosuusteemasta demokratiateemaiseksi.

Kehittämämme luokitus antaa tavan havainnollistaa ja määrällistää aineistolähtöisen LDA-mallin tulkintaan liittyviä hankaluuksia. Ihmisluokittelijoiden arviot LDA-aiheiden laadusta olivat samansuuntaisia C_v-koherenssilukuihin (Röder ym. 2015) perustuvat. Olemme suomentaneet aihemallinnuskäsitteistöä edelleen, kuten erottelun yksi- ja moniaihemalleihin, mutta kehittyvän metodiikan sanaston täsmentämisessä ja laajentamisessa on edelleen työtä tehtävänä. Uusin suomalainen tutkimus on tehnyt arvokkaan tienavauksen aihemallinnuksen soveltamiselle (Ahonen ja Wiberg 2018; Laaksonen ja Nelimarkka 2018; Purhonen ja Toikka 2016; Ylä-Anttila ym. 2018; Nelimarkka 2019). Sen ansiosta aihemallien käyttäjä voi toimia keskittyen politologiseen analyysiin eikä niukkaa julkaisutilaa tarvitse käyttää metodiikan oikeutukseen (Ahonen 2018). Aihemallinnusta käsittelevät review-artikkelit (Maier ym. 2018, Boumans ja

Trilling 2016, Grimmer ja Stewart 2013) antavat suuntaviivat metodin soveltamiselle. Monialainen yhteistyö ja tietojenkäsittelyosaaminen ovat käytännössä yhä edellytyksenä isojen aineistojen hallinnalle ja analyysille.

Laskennallisella tekstianalyysilla on paljon annettavaa politiikan tutkimukselle. LDA on yksi monista aihemalleista, ja automatisoidun sisällönanalyysin mahdollisuudet ovat huimat. Laskennallinen tutkimus ei kuitenkaan tee ihmistutkijaa tarpeettomaksi, vaan mallien tuottaman informaation laadun arviointi on vielä automatisoimatta sen merkitysten ymmärtämisestä puhumattakaan.

VIITTEET

1. Ahonen ja Wiberg (2018) kirjoittivat aihemalleista, ja esittivät myös englannin topic model -käsitteelle vaihtoehtoisen suomennoksen *topiikkamalli*. Laaksonen ja Nelimarkka (2018) käyttivät sanaa aihemalli hieman hämäävästi synonyymina LDA-aihemallille, vaikka erilaisia aihemalleja on lukuisia.
2. Samankaltaisiin tutkimuskysymyksiin voidaan vastata myös rakenteellisen aihemallintamisen avulla (Structural Topic Modeling). Rakenteelliset aihemallit käyttävät estimoinnissa metadataa, kun LDA perustuu yksinomaan dokumentin sisältämiin sanoihin.
3. Merkitsemme aineistossa esiintyvää sanaa tms. merkkijonoa ’puolilainauksin’.
4. LDA-mallissa on mahdollista säätää hyperparametreja α ja β, joilla on mahdollista hyödyntää ennakkotietoa datan priorijakaumasta (Wallach ym. 2009). Karkeasti yleistäen voitaneen sanoa, että pienemmät α-arvot toimivat oletuksena, että puheet koostuvat muutamasta dominoivasta aiheesta. Vastaavasti pienemmät β-arvot toimivat oletuksena siitä, että aiheet koostuvat muutamasta dominoivasta sanasta. Silloin esimerkiksi koulutuspolitiikalle tyypilliset sanat ’koulutus’, ’nuori’, ’yliopisto’, ’opiskelija’ ja ’koulu’ saisivat kyseisessä aiheessa suuren painon muihin sanoihin verrattuna. Käyttämämme mallit on laskettu Gensim-ohjelmakirjaston automaattisesti aineiston perusteella estimoitavalla epäsymmetrisellä α:lla (Řehůřek 2017), joka alustavissa mallinnuksissa tuotti ihmislukijalle huomattavasti helpommin hahmottuvia malleja kuin symmetrinen α eli ns. vanilja-LDA.
5. LDA-mallien laskennallisesta vertailusta ks. Maier ym. 2018, 102–103.
6. Puhemies ja varapuhemiehet pitävät vain keskustelua ja valtiopäivätoimia ohjaavia puheenvuoroja ottamatta kantaa käsiteltävään asiaan, poikkeuksena varapuhemies Jukka Mikkolan kuntien yleisestä kalleusluokituksesta annetun lain 6 §:n muuttamiseen liittyvä puheenvuoro (PTK 174/2002 vp), joka on sisällytetty aineistoon.
7. Avoimen lähdekoodin ohjelmisto on saatavilla osoitteessa http://turkunlp.github.io/Finnish-dep-parser/
8. Tässä yhteydessä emme käytä morfologian käsitettä lemma, koska jäsenninohjelma voi tuottaa putkimerkillä yhdistettyjä sanoja, joita ei esiinny luonnollisessa kielessä.
9. Tavallisesti LDA toimii hyvin riippumatta siitä, missä järjestyksessä algoritmi lukee puheet, vaikka laskennassa resursseista riippuen usein käsitellään vain osaa aineistosta kerrallaan (esim. chunksize=2000, jolloin ohjelma käsittelee kahtatuhatta tekstiä kerrallaan). Budjettikeskustelun kannatus-

puheenvuorot kuitenkin heikensivät mallien laatua, mikä ratkaistiin syöttämällä puheet ohjelmalle satunnaisessa järjestyksessä.

10. Koodaustyöstä kiitokset ansaitsevat Maria Anttila, Paul Hermansson ja Juuso Marttila.
11. Krippendorffin alfa on koodaajien yhdenmukaisuuden mittaluku, joka saa arvon yksi, kun kaikki koodaajat luokittelevat joka kohdan samoin, arvon nolla, kun luokitus vastaa satunnaista, ja negatiivisen arvon, kun yhdenmukaisuus on heikompi kuin satunnainen. Monista mittaluvuista tähän analyysiin valittiin Krippendorffin alfa siksi, että se on yleisesti käytetty ja toimii järjestysasteikollisille muuttujille. Mittaluku sopii myös välimatka- ja osuusasteikoille sekä nimellismuuttujille.

LÄHTEET

Ahonen, Pertti. 2018. Laskennalliset koneoppimisen menetelmät politiikan tutkimuksen kannalta: Institutionaalinen tarkastelu ja tieteenfilosofisen ja teoreettisen syventämisen tarve. *Politiikka* 60:2, 157–163.

Ahonen, Pertti ja Wiberg, Matti. 2018. Laskennallisten menetelmien mahdollisuuksia politiikan tutkimuksessa. *Politiikka* 60:1, 38–46.

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

Bayard, Pierre. 2009. *Miten puhua kirjoista joita ei ole lukenut*. 2. painos. Atena, Jyväskylä.

Belford, Mark, Namee, Brian ja Greene, Derek. 2018. Stability of topic modeling via matrix factorization. *Expert Systems with Applications*. Saatavilla: http://derekgreene.com/papers/belford18eswa.pdf Luettu: 19.9.2017.

Blei, David M., Ng, Andrew Y., Jordan, Michael I. ja Lafferty, John. 2003. Latent Dirichlet Allocation. *Journal Of Machine Learning Research* 3:4/5, 993–1022.

Blei, David M. ja Lafferty, John D. 2007. A correlated topic model of Science. *Annals of Applied Statistics*. 1:1, 17–35.

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

Boumans, Jelle W. ja Trilling, Damian. 2016. Taking Stock of the Toolkit. *Digital Journalism* 4:1, 8–23.

Brier, Alan, De Giorgi, Elisabetta ja Hopp, Bruno. 2016. Strategies in Computer-Assisted Text Analysis. *National Centre for Research Methods Working Paper* 03/16. Saatavilla: http://eprints.ncrm.ac.uk/3886/ Luettu: 14.11.2017.

Chuang, Jason, Gupta, Sonal, Manning, Christopher D. ja Heer, Jeffrey. 2013. *Topic Model Diagnostics: Assessing Domain Relevance via Topical Alignment*. International Conference on Machine Learning (ICML). Saatavilla: http://vis.stanford.edu/files/2013-TopicModelDiagnostics-ICML.pdf Luettu: 16.3.2017.

Curran, B., Higham, K., Ortiz, E. ja Vasques Filho, D. 2018. Look who's talking: Two-mode networks as representations of a topic model of New Zealand parliamentary speeches. *PLoS One* 13:6, e0199072.

Eduskunnan työjärjestys 17.12.1999/40 v. 2000. Finlex. Saatavissa: https://www.finlex.fi/fi/laki/ajantasa/2000/20000040 Luettu: 17.1.2018.

Greene, Derek ja Cross, James P. 2017. Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis* 25:1, 77–94.

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

Günther, Elisabeth ja Quandt, Thorsten. 2016. Word Counts and Topic Models.

Digital Journalism 4:1, 75–88.

Jelodar, Hamed, Wang, Yongli, Yuan, Chi, Feng, Xia, Jiang, Xiahui, Li, Yanchao ja Zhao, Liang. 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications* 78:11, 15169–15211.

Kleynhans, Neil. 2014. *Unsupervised Topic Modelling on South African Parliament Audio Data*. Saatavilla: https://researchspace.csir.co.za/dspace/bitstream/handle/10204/7947/Kleynhans3_2014.pdf Luettu: 14.11.2017.

Kontula, Anna. 2017. Politiikka ei avaudu laskimella. *Politiikka* 59:4, 298–300.

Kynčlová, Petra, Hron, Karel ja Filzmoser, Peter. 2017. Correlation Between Compositional Parts Based on Symmetric Balances. *Mathematical Geosciences* 49:6, 777–796.

Laaksonen, Salla-Maaria ja Nelimarkka, Matti. 2018. Omat ja muiden aiheet: Laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta. *Politiikka* 60:2, 132–147.

Liu, Shuhua ja Jansson, Patrik. 2017. City Event Identification from Instagram Data using Word Embedding and Topic Model Visualization. *Arcada Working Papers 7/2017*. Saatavilla: http://dspace.arcada.fi:8080/xmlui/handle/123456789/55 Luettu: 15.3.2018.

Maier, Daniel, Waldherr, A., Miltner, P., Wiedemann, G., Niekler, A., Keinert, A., Pfetsch, B., Heyer, G., Reber, U., Häussler, T., Schmid-Petri, H. ja Adam S. 2018. Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology. *Communication Methods and Measures*, 12:2-3, 93–118.

Mimno, David, Wallach, Hanna M., Talley, Edmund, Leenders, Miriam ja McCallum, Andrew. 2011. Optimizing semantic coherence in topic models. *In Proc. of the Conf. on Empirical Methods in Natural Language Processing*, 262–272.

Nelimarkka, Matti. 2019. Aihemallinnus sekä muut ohjaamattomat koneoppimismenetelmät yhteiskuntatieteellisessä tutkimuksessa: kriittisiä havaintoja. *Politiikka* 61:1, s. 6–33.

Nikolenko, Sergey I., Koltcov, Sergei ja Koltsova, Olessia. 2017. Topic Modelling for Qualitative Studies. *Journal of Information Science* 43:1, 88–102.

Palonen, Kari. 2012. *Parlamentarismi retorisena politiikkana*. Vastapaino, Tampere.

Pawlowsky-Glahn, Vera, Egozcue, Juan Jose ja Tolosana-Delgado, Raimon. 2015. *Modeling and analysis of compositional data*. Wiley, Chichester.

Pekonen, Kyösti. 2011. *Puhe eduskunnassa*. Vastapaino, Tampere.

Purhonen, Semi ja Toikka, Arho. 2016. Big datan haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät. Esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935–2015. *Sosiologia* 53:1, 6–26.

Quinn, Kevin M., Monroe, Burt L., Colaresi, Michael, Crepin, Michael H. ja Radev, Dragomir R. 2010. How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science* 54:1, 209–228.

Řehůřek, Radim. 2017. *models.ldamodel – Latent Dirichlet Allocation*. [ONLINE] https://radimrehurek.com/gensim/models/ldamodel.html Luettu: 4.10.2017.

Röder, Michael, Both, Andreas ja Hinneburg, Alexander. 2015. Exploring the Space of Topic Coherence Measures. *Proceeding WSDM '15 Proceedings of the Eighth ACM International Conference on Web*

Search and Data Mining, 399-408.
Sakamoto, Takuto ja Takikawa, Hiroki. 2017. *Cross-National Measurement of Polarization in Political Discourse. Analyzing floor debate in the U.S. and the Japanese legislatures*. Saatavilla: https://arxiv.org/pdf/1711.02977 Luettu: 14.11.2017.
Schofield, Alexandra, Magnusson, Måns ja Mimno, David. 2017. *Pulling Out the Stops: Rethinking Stopword Removal for Topic Models*. EACL 2017. Saatavilla: https://mimno.infosci.cornell.edu/papers/schofield_eacl_2017.pdf Luettu: 6.4.2017.
Sievert, Carson ja Shirley, Kenneth. E. 2014. LDAvis: A method for visualizing and interpreting topics. *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, 63–70.
Voutilainen, Eero. 2016. Tekstilajitietoista kielenhuoltoa: puheen esittäminen kirjoitettuna eduskunnan pöytäkirjoissa. Teoksessa: Tiittula, Liisa ja Nuolijärvi, Pirkko (toim.) *Puheesta tekstiksi. Puheen kirjallisen esittämisen alueita, keinoja ja rajoja*. Helsinki: SKS.
Wallach, Hanna M., Mimno, David M. ja McCallum, Andrew. 2009. *Rethinking LDA: Why priors matter*. Saatavilla: http://dirichlet.net/pdf/wallach09rethinking.pdf Luettu: 11.5.2017.
Winter, Lasse ja Wiberg, Matti. 2016. Lainsäädäntövolyymin kvantitatiivinen tekstintutkimus. *Edilex* 2016/34.
Ylä-Anttila, Tuukka, Eranti, Veikko ja Kukkonen, Anna. 2018. Aihemallinnuksesta kehysmallinnukseen. *Politiikka* 60:2, 148–156.

KIRJOITTAJATIEDOT

PETRI LOUKASMÄKI
Fil. yo., opiskelija
Tulevaisuuden teknologioiden laitos / Tietojenkäsittelyoppi
Turun yliopisto
lode@iki.fi

KIMMO MAKKONEN
VTM, tohtorikoulutettava
Filosofian, poliittisen historian ja valtio-opin laitos / Valtio-oppi
Turun Yliopisto
kimmo.makkonen@utu.fi