Erot sukupuolten välillä politiikan uutisoinnissa

Koneoppimista hyödyntävä analyysi

Iidaliisa Pardalin

Pro gradu –tutkielma

Kieliasiantuntijuuden tutkinto-ohjelma, digitaalinen kielentutkimus
Kieli- ja käännöstieteiden laitos
Humanistinen tiedekunta
Turun yliopisto

Marraskuu 2022

Pro gradu -tutkielma

Kieliasiantuntijuuden tutkinto-ohjelma, digitaalinen kielentutkimus
Iidaliisa Pardalin
Erot sukupuolten välillä politiikan uutisoinnissa – Koneoppimista hyödyntävä analyysi
Sivumäärät: 59 sivua, liitteet 9 sivua

Tämä tutkimus selvittää, onko sukupuolten välisiä eroja politiikan uutisoinnissa ja jos on, millaisia erot ovat. Aineistona on Yle Uutisten ja Helsingin Sanomien internet-sivuilla 31.12.2019–31.12.2021 julkaistuja uutisia. Aineistoa on tutkittu koneoppimismenetelmin korpusavusteisen diskurssianalyysin keinoja käyttäen. 53 167 Yle Uutisten ja 96 513 Helsingin Sanomien uutista käsiteltiin niin, että lopulliseen aineistoon saatiin Yle Uutisten osalta 16 659 ja Helsingin Sanomien osalta 17 023 virkettä, joissa mainitaan kansanedustaja joko koko nimeltä tai sukunimeltä. Näistä virkkeistä luotiin korpus, jonka avulla tehtiin laadullista analyysiä.

Tutkimusta vartien on ohjelmoitu tukivektorikoneluokittelija, joka pyrkii ennustamaan, käsitteleekö virke nais- vai mieskansanedustajaa. Sekä Yle Uutisten että Helsingin Sanomien aineiston osalta luokittelija kykeni perustasoa parempaan täsmällisyyteen; Yle Uutisten osalta perustaso oli 0.58 ja mallin täsmällisyys 0.67, Helsingin Sanomien perustaso oli 0.56 ja mallin täsmällisyys myös 0.67.

Analysoimalla luokittelijan luokittelussa käyttämiä piirteitä eli sanoja tutkimuksessa saatiin selville, millaisia eroavaisuuksia nais- ja mieskansanedustajia käsittelevissä uutisissa oli. Analyysiä tehtiin jakamalla sanoja eri ryhmiin sekä tekemällä virkkeistä luodun korpuksen avulla konkordanssitarkastelua siitä, millaisissa yhteyksissä kyseiset sanat esiintyivät alkuperäisissä virkkeissä.

Osa eroista on selitettävissä nais- ja mieskansanedustajien erilaisilla tehtävillä ja nimikkeillä, mutta muitakin eroja löytyi. Tutkimustulokset antava siis viitteitä siitä, että nais- ja mieskansanedustajien kohtelu mediassa on erilaista: esimerkiksi sanat sukupuoli, nainen ja äiti osoittautuivat erotteleviksi piirteiksi naisia käsittelevissä virkkeissä.

Laajempi aineisto, erilaiset painotukset sekä pyrkimykset selvittää mahdollisten erojen syitä olisivat tärkeitä jatkotutkimuksen aiheita.

Avainsanat: journalismi, koneoppiminen, politiikka, sukupuoli, tietokonelingvistiikka, korpusavusteinen diskurssianalyysi

Sisällysluettelo

1 Johdanto

Pääministeri Sanna Marinista tehtiin toukokuussa 2021 Seura-lehteen haastattelu (Lundelin, 2021). Artikkelin alussa kerrotaan, miten Marin kieltäytyy istumasta Kesärannan mättäälle kuvaa varten. Seuraavaksi tekstissä on Marinin kommentti ”Pyydettäisiinkö minulta tällaisia asioita, jos olisin miespääministeri?”, minkä jälkeen toimittaja epäilee, että näin ei luultavasti tehtäisi.

On pitkälti kiistatonta, että medialla on valtaa ja media muokkaa ihmisten käsityksiä eri aiheista. Se, miten jotakin aihetta tai henkilöä käsitellään mediassa, on siksi kiinnostavaa. Mikäli esimerkiksi naispoliitikkoja käsiteltäisiin miehiä negatiivisemmin, saattaisi tämä vaikuttaa äänestystuloksiin, sillä Suomessa äänestäjät seuraavat politiikkaa pääasiassa median kautta.

Eri sukupuolia olevien poliitikkojen mediassa saama kohtelu herättää paljon keskustelua. Kohdellaanko todella politiikassa vaikuttavia naisia mediassa ankarammalla kädellä kuin miehiä? Kysytäänkö ainoastaan naispoliitikoilta perheen ja uran yhdistämisestä? Tutkimustietoa asiasta ei Suomen kontekstissa ole. Suomalainen uutismedia seuraa itse lähinnä naisten ja miesten näkyvyyttä, ja moni media pyrkii näkyvyydessä sukupuolten väliseen tasa-arvoon (Kajander, 2021; Vasantola, 2021). Tutkimustulokset maailmalta antavat viitteitä siitä, että eroja voi hyvinkin löytyä (van der Pas & Aaldering, 2020).

Monilla poliitikoilla on Suomessa se käsitys, ettei media kohtele sukupuolia tasavertaisesti. Elina Talvitie haastatteli kirjaansa Keitäs tyttö kahvia: Naisia politiikan portailla (Talvitie, 2014) varten useita suomalaisia naispoliitikkoja, jotka kertoivat, että etenkin ulkoiseen olemukseen suhtautumisessa on eroja. Talvitie kirjoittaa, että miestenkin sojottaviin hiuksiin saatetaan kiinnittää huomiota, mutta miesten osalta tällaisia ulkonäköasioita ei yhdistetä poliittiseen kyvykkyyteen tai miehen uskottavuuteen, mitä taas naisten kohdalla tapahtuu. Suomalaista journalismin tekoa ohjaavat Julkisen sanan neuvoston (JSN) Journalistin ohjeet, joiden 26. kohta kuuluu seuraavasti:

> ”Jokaisen ihmisarvoa on kunnioitettava. Etnistä alkuperää, kansallisuutta, sukupuolta, seksuaalista suuntautumista, vakaumusta tai näihin verrattavaa ominaisuutta ei pidä tuoda esiin asiaankuulumattomasti tai halventavasti.”
>
> (Julkisen sanan neuvosto, ei pvm.)

Edellisen kerran JSN on ottanut kantaa tähän ohjeeseen ja poliitikosta uutisointiin liittyvään kanteluun syyskuussa 2022, kun JSN käsitteli Seiska-lehden nettisivuilla julkaistuun, pääministeri Sanna Marinista kertovaan juttuun, jossa käsiteltiin kuvien kera Marinin kehoa[1]. Päätöksessään JSN (2022) piti juttua naista esineellistävänä, mutta antoi Seiskalle vapauttavan päätöksen äänestystuloksella 7–6. Päätöksen mukaan Marinin ihmisarvoa ei loukattu ja siten hänen sukupuoltaan ei tuotu jutussa esiin 26. ohjetta rikkovalla tavalla.

Terveyden ja hyvinvoinnin laitos (THL) toteutti Sukupuolella väliä? -selvityksen, jonka osana selvitettiin sukupuolin roolia myös eduskunnassa (Siukola ym., 2020). Selvityksessä todetaan, että Suomessa saavutettiin tasapainoinen edustus (naisia ja miehiä 40–60 prosenttia kansanedustajista) vuonna 2007. Tutkimustulokset antavat viitteitä siitä, että media voi olla mukana vaikuttamassa myös tähän lukuun: vuonna 2019 julkaistu, yli 50 maata (ml. Suomi) kattanut tutkimus totesi, että median seksistisyyden ja naisehdokkaiden välillä on negatiivinen yhteys (Haraldsson & Wängnerud, 2019). Mitä seksistisempi media siis on, sitä vähemmän naisehdokkaita oli. Arvioidessaan sukupuolen ja median tutkimuksen tärkeimpiä suuntauksia Byerly (2012) nostaa esille juuri representaation yhtenä tärkeimmistä tutkimussuuntauksista, ja toteaa, että myös uutisten naiskuva on tiukasti kiinni patriarkaalisessa naiskäsityksessä. Representaatio ei kuitenkaan koske pelkästään naisia, vaan on tärkeää tutkia myös sitä, miten miehiä kuvataan. Silläkin on väliä, minkälaisen ”miehen mallin” media antaa. Tämä tutkimus ei siis keskity siihen, minkälaisen naiskuvan uutisointi maalaa, vaan selvittää sitä, onko uutisoinnissa naisten ja miesten välillä eroja ja jos on, millaisia erot ovat. Tähän syvennytään käyttämällä digitaalisen kielentutkimuksen keinoja; koneoppimista ja korpusavusteista diskurssianalyysiä. Aiheensa puolesta tutkimus linkittyy vahvasti journalismin tutkimuksen sekä myös sukupuolentutkimuksen viitekehyksiin.

1.1 Tutkimuskysymykset

Tämän tutkimuksen tutkimuskysymykset siis ovat:

1. Löytyykö politiikan uutisoinnista naisia ja miehiä koskevissa uutisista eroavaisuuksia?

[1] Seiska 4.6.2022: Paparazzi iski: kireät trikoot myötäilivät Sanna Marinin huipputreenattua peppua – kuva! https://www.seiska.fi/Kotimaa/Paparazzi-iski-kireat-trikoot-myotailivat-Sanna-Marinin-huipputreenattua-peppua-kuva (linkki tarkastettu 1.11.2022)

2. Millaisia erot ovat?

Tutkimuksen aineistona on Yle Uutisten ja Helsingin Sanomien internet-sivuilla 31.12.2019–31.12.2021 julkaistuja uutisia. Yleisradion uutiset eivät ole kaupallisen median tuottamaa, joten sisällöt ovat kaikkien luettavissa eikä aineisto siten sisällä maksumuurin takana olevia uutisia. Helsingin Sanomat puolestaan on kaupallinen media, mutta aineistossa ei ole mukana maksumuurin takana olleita uutisia. Uutiset on poimittu RSS-syötteestä.

Aineistoa tutkitaan koneoppimismenetelmin. Digitaalisen kielentutkimuksen keinoja käyttäen voidaan tutkia erittäin laajoja, kymmeniä tuhansia uutisartikkeleita sisältäviä aineistoja, joiden tutkiminen perinteisin diskurssianalyysin keinoin olisi mahdotonta toteuttaa. Koneoppimismenetelmien avulla laajasta aineistosta voidaan poimia tämän tutkimuksen kannalta relevantit uutisartikkelit, eli ne, jotka käsittelevät poliitikkoja. Uutiset on siis käyty koneellisesti läpi niin, että mukaan on otettu ainoastaan sellaiset uutiset, joissa on mainittu kansanedustaja nimeltä. Jakoa naisista ja miehistä kertoviin artikkeleihin ei myöskään tarvitse tehdä käsin, vaan sen voi tehdä automatisoidusti. Näistä uutisista on poimittu lopulliseen korpukseen ne virkkeet, joissa kansanedustaja mainitaan joko koko nimeltä tai sukunimeltä. Tutkimusta varten on ohjelmoitu luokittelija, jonka koulutusaineistona on nais- ja mieskansanedustajia käsittelevien uutisartikkeleiden virkkeitä. Malli pyrkii koulutusdatan perusteella ennustamaan ennen näkemättömästä tekstistä, käsitelläänkö siinä nais- vai mieskansanedustajaa.

Mallia tarkastelemalla voidaan tehdä päätelmiä siitä, millä perusteella malli tekee ennustuksensa. Näitä kielellisiä piirteitä tutkimalla voidaan löytää piirteitä, jotka ovat todennäköisemmin tietyn sukupuolen edustajaa käsittelevässä artikkelissa. Samanlaista tutkimusmenetelmää on käyttänyt esimerkiksi Leavy (2019) tutkiessaan politiikan uutisointia Irlannissa ja siinä esiintyviä naisten ja miesten eroja. Lähestymistavassa sovelletaan korpusavusteisen diskurssianalyysin keinoja (*corpus-assisted discourse analysis*, lyh. CADS, kts. esim. Partington ym., 2013), eli aineistoa lähestytään ensin kvantitatiivisesti ja sen jälkeen kvalitatiivisesti tehden tulkintoja erojen laadusta. Sitä, minkälaisissa yhteyksissä tai millaista kieltä käyttäen miehistä ja naisista suomalaisessa mediassa uutisoidaan, ei ole vastaavanlaista laajaa aineistoa ja koneoppimista käyttäen Suomessa ennen tätä tutkittu.

Tekstissä käydään ensin luvussa 2 läpi sitä, mitä sukupuolittunut kieli on ja tarkastellaan lyhyesti suomen kielen sukupuolittuneita piirteitä. Luvussa 3 käyn läpi, miten naisten ja

miesten välisiä eroja on tutkittu keskittyen politiikan uutisointiin. Tämän jälkeen luvussa 4 esittelen tutkimukseen valittuja menetelmiä ja selitän lyhyesti koneoppimista, jotta lukijalla on käsitys siitä, miten tutkimukseen käytetyt metodit toimivat. Luku 5 keskittyy aineistoon, kuvaillen vaiheet, jotka aineiston käsittelyssä käytiin läpi sekä esitellen aineiston yleispiirteitä. Varsinainen analyysi on luvussa 6, minkä jälkeen luvussa 7 käydään läpi tutkimuksen johtopäätökset sekä pohditaan niitä kysymyksiä, joihin tämän tutkimuksen puitteissa ei pystytty vastaamaan sekä esitetään vaihtoehtoja mahdolliselle jatkotutkimukselle.

1.2 Kielestä, terminologiasta ja sukupuolen käsitteestä

Tämä tutkimus käsittelee suomalaista ja suomenkielistä uutisointia, ja siksi myös tutkimuksen esityskieleksi on valittu suomi.

Digitaalinen kielentutkimus on vahvasti kansainvälinen tutkimusala, ja suuri osa tutkimuksesta julkaistaan englanniksi. Osin tästä syystä ja osin tutkimusalan alati muuttuvasta ja nopeasti kehittyvästä luonteesta johtuen kaikille termeille ei välttämättä ole vakiintunutta suomenkielistä käännöstä. Jotta lukija tietää varmasti, mistä käsitteestä on kyse ja voi helposti yhdistää lukemaansa muuhun englanninkieliseen tutkimukseen, on termejä selittäessä niistä annettu myös niiden englanninkielinen vastine.

Tutkimuksessa keskitytään naisiin ja miehiin, vaikka nykytiedon mukaan sukupuolia on enemmän kuin nämä kaksi. Tutkimuksen keskittyminen nainen-mies-dikotomiaan on perusteltua siksi, että tutkimuksen tekohetkellä Suomessa juridisia sukupuolia ei ole enempää. Eduskunta julkaisee tietoa kansanedustajien sukupuolijakaumasta huomioiden vain nämä kaksi sukupuolta. On mahdollista, että joku aineistossa naiseksi tai mieheksi katsottu poliitikko identifioituu muuksi kuin hänelle merkityksi sukupuoleksi. Tätä ei ole otettu tutkimuksessa huomioon.

1.3 Eettiset ja tekijänoikeudelliset näkökulmat

Tämän tutkimuksen aineisto ei käsittele esimerkiksi henkilötietoja tai muita arkaluontoisia tietoja. Tutkimus on toteutettu Tutkimuseettisen neuvottelukunnan (TENK) ohjeita noudattaen.

Vaikka aineiston uutiset ovat olleet vapaasti saatavilla internetissä, ne ovat tekijänoikeuksien alaisia. Opetus- ja kulttuuriministeriö on kuitenkin 2012 hankkinut Kopiostolta digiluvan,

jonka nojalla avoimilta verkkosivuilta voi skannata ja kopioida materiaalia tutkimuskäyttöön. Uutisten käyttäminen tällä tavalla tutkimusmateriaalina ei siis loukkaa tekijänoikeuksia.

2 Sukupuolittunut tai seksistinen kieli

Suomen kielessä ole kieliopillista sukua (vrt. esim. saksan maskuliini, feminiini ja neutri *der / die / das*) ja suomen kielessä on sukupuolineutraali persoonapronomini hän (vrt. esim. englannin *he / she*). Tarkempi tarkastelu ja tutkimus kuitenkin osoittaa, että suomen kielessä on useita sukupuolittuneita aspekteja, vaikka kieli onkin kieliopillisesti sukupuolineutraali. Käyn tässä kappaleessa läpi lyhyesti suomen kielen sukupuolittuneisuutta ja seksististä kielenkäyttöä myöhemmän laadullisen analyysin tueksi ja taustaksi. Sukupuolittunutta ja seksististä kielenkäyttöä yleisemmin on tutkittu laajalti 1960-luvulta lähtien (Mills, 2008), mutta keskityn nyt erityisesti suomen kieleen.

Suomen kielen sukupuolittuneisuutta ja seksismiä on tutkinut etenkin Mila Engelberg, joka on käsitellyt aihetta esimerkiksi väitöskirjassaan Yleispätevä mies – Suomen kielen geneerinen, piilevä ja kieliopillistuva maskuliinisuus (Engelberg, 2016) sekä osittain siihen perustuvassa teoksessa Miehiä ja naisihmisiä – Suomen kielen seksismi ja sen purkaminen (Engelberg, 2018). Kielen sukupuolittuneisuutta käsitellessä onkin tärkeä erottaa niin sanottu neutraali sukupuolittuminen seksismistä. Esimerkiksi sanat mies ja nainen ovat luonnollisesti sukupuolittuneita sanoja. Engelberg kirjoittaa aiemmin mainitussa teoksessaan (Engelberg, 2018) seuraavasti: ”Seksistisen kielenkäytön reformin kohteena ei ole ollut kielellisesti ilmaistu sukupuoli sinänsä, vaan kielen käyttö sukupuolten eriarvoistamisen välineenä”. Päällisin puolin neutraaleja sanoja, kuten mainitut mies ja nainen tai tyttö ja poika, voi siis myös käyttää seksistisesti (esimerkiksi tytöttely, tai feminiinisten sanojen käyttäminen heikkouden tai huonouden ilmaisuna, kuten ”heittää kuin tyttö”).

Suomen kielessä pääte -mies on hyvin yleinen erilaisissa ammattinimikkeissä, kuten esimerkiksi puhemies, lakimies, palomies, tiedemies, esimies, virkamies ja niin edelleen. Virallisissa yhteyksissä tällaisia sukupuolittuneita titteleitä on alettu joillain aloilla ja joissain yhteyksissä vaihtaa sukupuolineutraaleiksi: näin ovat tehneet esimerkiksi THL (Siukola & Teräsaho, 2021), Aamulehti (2019) ja Duunitori (Duunitori, n.d.). Tutkimusten mukaan -mies-päätteelisiä ammattinimikkeitä ei nähdä täysin sukupuolineutraaleina. Engelberg kirjoittaa esimerkiksi tutkimuksesta, jossa koehenkilöille esitettiin tekstissä henkilö virkamies-tittelillä, ja pyydettiin keksimään henkilölle nimi ja piirtämään hänen kuvansa. 80 prosenttia määritteli henkilön mieheksi (Engelberg, 2000).

Englannin kielessä sanalla *man* voidaan tarkoittaa yleisesti ihmisiä, ja myös suomessa on samanlaisia piirteitä. Kuvaava esimerkki tästä ilmiöstä on termi jokamiehenoikeudet, jolla viitataan oikeuksiin, jotka koskevat kaikkia ihmisiä sukupuolesta riippumatta. Miessukupuoli on siis usein kielen käytössä niin sanottu oletussukupuoli: ilmiötä selittää kielen yleispätevä maskuliinisuus, joka on piirteenä monissa kielissä (Gygax ym., 2019). Suomen -mies-päätteiset ammattinimikkeet ovat osa kielen yleispätevää maskuliinisuutta: -mies-päätteisellä ammattinimikkeellä voidaan siis viitata mihin tahansa sukupuoleen. Esimerkkejä samasta ilmiöstä toisin päin on hyvin haastava löytää. Esimerkiksi lentoemäntä-titteliä ei käytetä miehistä paitsi halventavassa yhteydessä, miehistä (ja nykyään enenevissä määrin myös naisista) käytetään nimikettä stuertti.

Sukupuolen esiin tuominen liittyy myös kielen yleispätevään maskuliinisuuteen. Feministisessä teoriassa ilmiöstä puhutaan esimerkiksi nimellä MAN, male-as-norm (Engelberg, 2016). Miessukupuolisuus on oletus, ja tästä poikkeaminen tuodaan esille. Esimerkkejä tällaisesta kielenkäytöstä ovat esimerkiksi sanat naisurheilu, naispoliitikko tai naislääkäri. Monet sanat sisältävät niin sanotun piilosukupuolen, vaikka ne olisivat päällisin puolin sukupuolineutraaleja: Engelberg nostaa esimerkeiksi sanat sairaanhoitaja, prostituoitu, seppä ja vanki, jotka herättävät mielikuvan sukupuolesta, vaikka sanat itsessään ovat sen suhteen neutraaleja (Engelberg, 2018).

Kotimaisten kielten keskuksen ylläpitämästä Kielitoimiston sanakirjasta löytyy joitakin sanoja, joiden sukupuolittuneisuus on tuotu sanakirjamääritelmässä esille: niiden määritelmissä kerrotaan, että sanaa käytetään ainakin joissain yhteyksissä varsinkin miehistä tai naisista. Tällaisia ovat miessukupuolen osalta esimerkiksi hujoppi, irstailla ja raavas, naissukupuolen osalta puolestaan hemaiseva, muodokas ja emansipoitua (*Kielitoimiston Sanakirja*, ei pvm.). Samanlaisia sanoja on muitakin, esimerkiksi komea liitetään yleensä miehiin ja kaunis naisiin; kaunis mies ja komea nainen -yhdistelmillä on puolestaan omanlaisensa konnotaatiot.

Suomen kielen osalta laajaa tutkimusta siitä, mitkä sanat mielletään feminiinisiksi ja mitkä maskuliinisiksi, ei ole tehty. Monet päätelmät sanojen mahdollisista feminiinisistä tai maskuliinisista konnotaatioista perustuvat monissa diskurssianalyyseissäkin usein yleiseen käsitykseen sanojen mahdollisista sukupuolisidonnaisuuksista.

3 Sukupuolten väliset erot politiikkauutisoinnin tutkimuskohteena

Aiemman tutkimuksen siitä, miten sukupuolierot näkyvät politiikan uutisoinnissa, voi jakaa karkeasti kvantitatiiviseen ja kvalitatiiviseen. Naisten ja miesten medianäkyvyyden määrää on tutkittu paljon, ja monet mediat Suomessa seuraavat myös itse oman uutisointinsa sukupuolijakaumaa ja kertovat siitä avoimesti (Kajander, 2021; Vasantola, 2021). Kvalitatiivisessa tutkimuksessa syvennytään siihen, minkälaisia eroja naisista ja miehistä kertovassa uutisoinnissa on. Tässä luvussa esitän katsauksen siihen, miten aihetta on näistä kahdesta näkökulmasta tutkittu sekä käyn läpi menetelmiä, joilla medianäkyvyyttä on tutkittu.

3.1 Kvantitatiivinen näkökulma: medianäkyvyyden määrän tutkimus

Sukupuolten väliseen eroon kaikessa uutisoinnissa on alettu kiinnittää viime vuosina aiempaa aktiivisemmin huomiota, ja eroa voi jopa seurata lähes reaaliajassa the Gender Gap Tracker[2] -palvelun avulla internetissä. Suuri osa tutkimuksista ja laskelmista keskittyy yleensä uutisointiin laajemmin, eikä nimenomaan politiikan uutisointiin.

Oletus usein on, että uutisointi kuvastaa tosielämää: jos kansanedustajista 40 prosenttia on naisia, kansanedustajista kertovasta uutisoinnista 40 prosenttia kertoisi naiskansanedustajista. Uutisointia tekevät kuitenkin ihmiset, joilla voi olla implisiittisiä tai eksplisiittisiä ennakkoasenteita, jotka voivat vaikuttaa uutisointiin. Miehen haastateltavaksi valitseva toimittaja ei välttämättä tietoisesti tee valintaa olla ottamatta haastatteluun naista. Irlannissa parlamentin alahuoneen Dáil Éireannin naisehdokkaiden määrä kasvoi 90 prosenttia vuosien 2011 ja 2016 välillä (syynä oli lainsäädäntö, joka saneli puolueiden menettävän puolet julkisesta rahoituksestaan, mikäli naisia tai miehiä olisi ehdokkaana alle 30 prosenttia), mutta tutkimus osoitti sukupuolten välisen eron uutisoinnissa vain kasvavan entisestään naisehdokkaiden määrän kasvaessa (Courtney ym., 2020). Sama ilmiö näkyi tutkimuksessa myös urheilu-uutisoinnissa, joten kyseessä ei voida nähdä olevan politiikan uutisoinnin erityispiirre.

Suomessa uutisoitiin toukokuussa 2021 Retriever-media-analyysiyhtiön raportista, jossa kerrottiin vuonna 2020 politiikan uutisoinnissa naisten näkyneen miehiä enemmän:

[2] https://gendergaptracker.informedopinions.org/

mediamaininnoista naisten osuus oli 52 prosenttia, vaikka kansanedustajista naisia oli 46 prosenttia (Retriever, 2021). Merkittävää kuitenkin on, että vuonna 2020 hallitus sai koronatilanteen vuoksi erityisen paljon huomiota, ja pääministeri Sanna Marinin johtama hallitus oli naisvetoinen. Kun ministerien maininnat jätettiin analyysista pois, naiset saivat enää 29 prosenttia medianäkyvyydestä.

3.2 Kvalitatiivinen näkökulma: millaista medianäkyvyys on?

Ensimmäisten joukossa erojen laatua tutkivat Kahn ja Goldenberg (1991), jotka tulivat siihen tulokseen, että yhdysvaltalainen media keskittyi senaatinvaalien uutisoinnissa naisehdokkaiden kohdalla miehiä enemmän ehdokkaiden voittomahdollisuuksiin kuin heidän kantoihinsa poliittisissa kysymyksissä.

Tutkimusten tulokset vaihtelevat. Sukupuolieroja poliittisessa uutisoinnissa käsittelevässä meta-analyysissä (van der Pas & Aaldering, 2020) todettiin, että eri tutkimuksissa on päädytty hyvin eri tuloksiin: tutkimuksia, joissa naisia käsitellään positiivisemmin, oli analyysissa yhtä paljon kuin niitä, joissa miehiä käsiteltiin positiivisemmin sekä niitä, joissa eroja ei havaittu.

Suuri osa tutkimuksista seuraa Kahnin ja Goldbergin jalanjälkiä ja keskittyy vertailemaan medianäkyvyyseroja juuri kampanja-aikaan (esim. Aaldering & van der Pas, 2020), mutta myös pidempiaikaista tutkimusta löytyy. Esimerkiksi Leavy (2019) on tutkinut politiikan journalismin sukupuolista eriarvoistamista koneoppimisen keinoin pitkällä aikavälillä. Leavyn irlantilaista mediaa koskevassa tutkimuksessa naispoliitikot saivat osakseen tiukempaa kritiikkiä, ja heidän perhe-elämäänsä käsiteltiin miehiä useammin. Naisministereistä käytettiin myös miehiä useammin etunimeä. Miesministerien alkoholinkäytöstä puolestaan kerrottiin mediassa naisia useammin.

On tärkeää ottaa huomioon, ettei politiikan uutisointi ole erillään sitä ympäröivästä yhteiskunnasta ja sen stereotypioista. Alankomaalaisia uutisia vuosilta 2006–2012 analysoinut tutkimus (Aaldering & van der Pas, 2020) tutki sukupuolten välisiä eroja uutisoinnissa ja nosti esille erityisesti johtajuuteen liittyvät sukupuolittuneet stereotypiat ja niiden ottamisen huomioon yleisten sukupuolistereotypioiden lisäksi. Tutkimuksessa löydettiin eroja siinä, miten johtajuuteen liittyvistä piirteistä kirjoitettiin mediassa miesten osalta naisia useammin.

Humprecht ja Esser (2017) tutkivat naisten ja miesten välisiä eroja verkossa julkaistuissa politiikan uutisissa kuudessa eri maassa, Yhdysvalloissa, Isossa-Britanniassa, Saksassa,

Sveitsissä, Ranskassa ja Italiassa. Eroja löytyi esimerkiksi aiheissa, joissa naisia ja miehiä haastateltiin, sekä siinä, miten naiset ja miehet visuaalisesti esitettiin uutisoinnissa: naiset esiintyivät uutisiin liittyvissä kuvissa huomattavasti miehiä useammin. Aiheet uutisissa, joissa naisia ja miehiä haastateltiin, erosivat selkeästi ja naisista kirjoitettiin miehiä useammin kolumneja tai vastaavia ei-uutissisältöjä.

Toisenlaista näkökulmaa tarjoavat Hayes & Lawless (2016), joiden kirja Women on the run: Gender, media and political campaigns in a polarized era käsittelee asiaa Yhdysvalloista käsin. Hayes & Lawless toteavat, että vaikka stereotypioita tai ennakkoasenteita naisia kohtaan olisi ollut aiemmin, nykyaikana vastaavia ilmiöitä ei löydy ainakaan siinä määrin, että sillä olisi vaikutusta esimerkiksi kongressiedustajien sukupuolijakaumaan. Kirjassa esitellään analyysi paikallislehtien uutisoinnista vuosien 2010 ja 2014 välivaalien ajalta ja todetaan, ettei nais- ja miesehdokkaista kertovassa uutisoinnissa ole merkittäviä eroja. Sukupuolisyrjintään viittaavia eroja ei löytynyt myöskään Sveitsin vuoden 2015 parlamenttivaaleja edeltävää uutisointia analysoivassa tutkimuksessa (Rohrbach ym., 2020). Vaaleja edeltävää paikallisuutisointia Meksikossa analysoinut Vidal-Correa (2020) tuli siihen tulokseen, että uutisointi oli pitkälti neutraalia ja muut tekijät, kuten ehdokkaan voittomahdollisuudet, olivat sukupuolta tärkeämpiä: mitä todennäköisempänä media piti ehdokkaan voittoa, sitä enemmän hän näkyi uutisissa, eikä sukupuolella voitu osoittaa olevan vaikutusta asiaan.

Mahdollinen sukupuolisyrjintä tai muut sukupuolten väliset erot uutisoinnissa ovat aina ajasta ja paikasta riippuvaista. Yksi taustalla vaikuttava tekijä voi olla naispoliitikoiden määrä: Australian ja Kanadan naispääministereitä[3] koskevaa uutisointia analysoineet Trimble ym. (2021) toteavat, että tulosten valossa naisia koskevan uutisoinnin keskittyminen henkilöön vähenee sitä mukaa mitä enemmän naisia on poliittisissa johtotehtävissä. Kanadan provinssien pääministereistä kertovan uutisoinnin mahdollista sukupuolittuneisuutta tutkineet Thomas ym. (2021) löysivät eroja paitsi uutisoinnin määrästä (naisjohtoisista hallituksista uutisoitiin vähemmän), myös sen laadussa. Naisista kertovissa uutisissa puhuttiin esimerkiksi vaatteista useammin ja niissä käytettiin enemmän sukupuolittuneita ilmaisuja.

[3] Alueellisten hallintoalueiden pääministerit, engl. *premier*

Tutkimusta on tehty myös länsimaiden ulkopuolella. Globaalia naispoliitikoista kertovaa uutisointia yli 30 vuoden ajalta analysoinut tutkimus osoitti, että länsimainen uutisointi on keskittynyt naisten ulkonäköön huomattavasti afrikkalaista ja aasialaista journalismia enemmän, tosin länsimaissakin ulkonäöstä puhuva uutisointi on tutkimuksen mukaan vähentynyt 2000-luvulla (Joshi ym., 2020). Länsimaiden ulkopuolella naispoliitikot näkyivät usein hyveellisessä roolissa ja heitä kuvattiin vähemmän korruptoituneina. Mielenkiintoinen huomio tutkimuksessa oli myös se, että naispoliitikot nähtiin nimenomaan sukupuolensa kautta, eikä muita intersektionaalisia tekijöitä otettu yleensä huomioon. Naispoliitikot usein esitettiin patriarkaalisen järjestelmän uhreina riippumatta siitä, miten etuoikeutettu tausta naiseuden lisäksi poliitikolla saattoi olla.

Yksi erottava tekijä naisia ja miehiä käsittelevässä uutisoinnissa on ns. uutuudenviehätys (*novelty*): ensimmäiset roolissa vaikuttavat naiset saivat usein huomattavaa mediahuomiota (Thomas ym., 2021; Trimble ym., 2021). Vastaavaa ilmiötä ei ole olemassa miehillä, sillä poliittisia virkoja, joihin olisi viime vuosina tai vuosikymmeninä valittu ensimmäinen mies, ei ole.

Monet suomalaiset nais- ja miesrepresentaatiota politiikan journalismissa tutkivat artikkelit ja pro gradu -työt keskittyvät yleensä hyvin pieneen aineistoon. Tutkimuksia on tehty esimerkiksi siitä, miten sukupuolirepresentaatiot eroavat aikakauslehtien kansikuvissa, sivuten samalla sitä, miten mies- ja naispoliitikoista esitetyt kuvat eroavat (Luhtakallio, 2016). Heidi Salminen on tutkinut journalistiikan pro gradu -työssään politiikan journalismin sukupuolittumista Ylen vaalivideoissa (2018) ja Joonas Lehtonen puolestaan politiikan sukupuolittunutta työnjakoa Ylen televisiouutisissa (2018).

4 Tutkimusmenetelmät

Monet sukupuolten välisiä eroja käsittelevät tutkimukset keskittyvät, kuten edellisessä luvussa on kerrottu, melko pieniin aineistoihin. Diskurssianalyysin tekeminen laajoista aineistoista vie suuret määrät resursseja, joten tämä on ymmärrettävää. Koneoppimista käyttämällä voidaan kuitenkin analysoida suuria aineistomääriä uudella tavalla. Tämän tutkimuksen menetelmäksi on tästä syystä valittu koneoppiminen, tarkemmin ottaen tukivektorikoneluokittelija. Tämä mahdollistaa suuren, kymmeniä tuhansia uutisartikkeleita käsittelevän aineiston tutkimisen.

Kun puhutaan suurten tekstiaineistojen analysoinnista ja tutkimisesta, puhutaan yleensä korpuslingvistiikasta. Korpuslingvistiikan perinteisiä menetelmiä ovat esimerkiksi avainsana- ja kollokaatioanalyysi (Kyröläinen & Laippala, 2020). Kuten Kyröläinen ja Laippala toteavat, avainsana-analyysissä koko aineisto nähdään yhtenä kokonaisuutena, eikä se siten kuvaa eroja aineiston yksittäisten tekstien välillä. Kollokaatioanalyysissä puolestaan analysoidaan tiettyjen sanojen esiintymistä toistensa kanssa. Jotta tiettyjen sanojen kollokaatteja voitaisiin analysoida, pitäisi pystyä valitsemaan tutkimuksen kannalta relevantit sanat. Aineistosta voisi esimerkiksi etsiä kollokaatteja eri nais- ja mieskansanedustajien nimille, mutta koska kollokaatit ovat ainoastaan sanan lähistöllä, ts. korkeintaan muutamien sanojen päässä esiintyviä sanoja, tällainen analyysi voisi jättää paljastamatta monia tekstien erilaisia piirteitä. Kyröläinen ja Laippala nostavatkin esille tukivektorikoneen esimerkkinä kielentutkimuksessa usein käytetystä algoritmista (Kyröläinen & Laippala, 2020).

Käyn tässä kappaleessa lyhyesti läpi koneoppimisen tärkeimmät termit ja syvennyn sitten ennustavan ja selittävän mallinnuksen eroihin ja siihen, miksi tukivektorikoneluokittelija sopeutuu tutkimusmenetelmäksi tämän tutkimuksen tyyppisessä korpusavusteisessa diskurssianalyysissä. Lopuksi esittelen luokittelijan arvioinnin mittareita, joita käytän seuraavassa luvussa tulosten tulkitsemisessa.

4.1 Lyhyesti koneoppimisesta ja luokittelusta

Koneoppiminen (*machine learning*) on tekoälyn (*artificial intelligence, AI*) osa-alue. Sillä on vahva pohja tilastotieteessä ja todennäköisyyslaskennassa: monet koneoppimismallit käytännössä laskevat tilastollisia todennäköisyyksiä eri ilmiöille. Esimerkiksi koneoppimisen keinoja käyttäen opetettu kasvojentunnistusohjelma ei oikeastaan tunnista kasvoja, ainakaan sillä tavalla, miten ihminen tunnistamisen yleensä ymmärtää, vaan arvioi kenelle kasvot

kuuluvat suurimmalla todennäköisyydellä. Ohjelma ei myöskään varsinaisesti näe kasvoja: se käsittelee dataa.

Tärkeimpiä jaotteluja koneoppimisessa on jako ohjattuun ja ohjaamattomaan oppimiseen (*supervised vs. unsupervised learning*). Ohjatussa oppimisessa mallille syötetyssä opetusdatassa on jollakin tavalla merkitty opittava asia: jos on tarkoitus opettaa erottamaan omenan kuva, mallille on syötetty omenoiden (ja mahdollisesti muiden hedelmien) kuvia kertoen, mitkä niistä ovat omenia. Ohjaamattomassa oppimisessa ns. oikeita vastauksia ei ole annettu mallille. Nyt mallille ei siis kerrota, mitkä hedelmistä esittävät omenoita, vaan se saa ainoastaan erilaisten hedelmien kuvia. Malli voi ohjaamattomasti etsiä erilaisia klustereita aineistosta ja siten se saattaa päätyä erottelemaan omenat ja appelsiinit toisistaan.

Luokittelu (*classification*) on hyvin tyypillinen koneoppimisen tehtävä. Yleisesti käytetty esimerkki tekstin luokittelun kentältä on sähköpostin roskapostisuodatin: se luokittelee saapuvan sähköpostin joko roskapostiksi tai ei-roskapostiksi. Muita juuri kieleen ja tekstiin liittyviä luokitteluongelmia ovat esimerkiksi tunneanalyysi (*sentiment analysis*) tai topiikkimallinnus (*topic analysis*). Luokittelijan tärkeimpiä ominaisuuksia on sen yleistämiskyky (*generalization*), eli kuinka hyvin malli toimii ennennäkemättömän datan luokittelussa.

Luokittelua varten on luotu useita algoritmeja ja keinoja, joista tässä esitellään yleisimpiä.

Naiivi Bayesin luokitin (*naive Bayes classifier*) on Bayesin teoreemaan perustuva, todennäköisyyslaskentaa käyttävä luokitin. Naiivi Bayesin luokitin käyttää ohjattua oppimista. Koska luokitin olettaa piirteiden olevan riippumattomia toisistaan, ts. jokainen piirre vaikuttaa yhtä paljon siihen, kuuluuko syöte johonkin luokkaan tai ei, sitä sanotaan naiiviksi.

Tukivektorikone (*support vector machine, SVM*) (kts. esim Vapnik, 1995) on luokittelumalli, joka binäärisessä luokittelussa sovittaa n-ulotteisessa vektoriavaruudessa havaintojen välille sellaisen hypertason, joka erottaa kaikki eri luokkiin kuuluvat havainnot toisistaan mahdollisimman suurella marginaalilla. Marginaalin reunoille jääviä pisteitä kutsutaan tukivektoreiksi. Tässä mallissa syöte muunnetaan piirrevektoriksi (*feature vector*), joka on n-ulotteinen esitys syötteen piirteistä. Ulottuvuuksien määrä eli n on piirteiden määrä eli esimerkiksi eri sanojen määrä. Tukivektorikoneen etu esimerkiksi seuraavaksi esitettäviin

neuroverkkoihin verrattuna on sen selitettävyys: luokitteluun eniten vaikuttaneet piirteet on mahdollista saada hyvin yksinkertaisesti selville.

Neuroverkot (*neural networks*) on kehitetty mallintamaan ihmisaivojen toimintaa, ja ne ovat äärimmäisen monimutkaisia ja vaativat erittäin paljon laskentatehoa. Myös neuroverkkoja voidaan käyttää erilaisten luokittelujen tekemiseen, ja esimerkiksi tukivektorikone voidaan toteuttaa neuroverkkojen avulla. Monet neuroverkkosovellukset ovat niin sanottuja musta laatikko -algoritmeja (*black box algorithm*). Tämä tarkoittaa sitä, ettei neuroverkkojen kehittäjillekään ole täysin selvää, miten neuroverkko tekee esimerkiksi luokitteluennusteensa. Tästä syystä luokitteluun vaikuttaneiden piirteiden analysointi voi osoittautua vaikeaksi – joskus jopa mahdottomaksi. Teknologiat kehittyvät kuitenkin koko ajan ja mallien selittäminen ja analysointi onkin yksi koneoppimisen ja tekoälyn tärkeimpiä kehitys- ja tutkimussuuntauksia (kts. esim. Biecek & Burzykowski, 2021).

Luokittelu on ohjattua oppimista. Ohjatussa koneoppimisessa on erittäin tärkeää, että mallin koulutukseen tai sovittamiseen käytetty data on laadukasta. Ilman laadukasta dataa mallin yleistämiskyky jää todennäköisesti alhaiseksi. Mallin opettamisessa käytetään yleensä kolmeen osaan jaettua dataa: opetus-, validointi- ja testidataa (*training, validation / development, test data*). Opetusdatalla malli opetetaan, validointidataa käytetään hienosäätöön ja testidatan avulla testataan, miten hyvin malli opetetun pohjalta soveltuu ennennäkemättömään dataan. Testidata on siis dataa, jota malli ei ole ennen nähnyt, mutta josta oikeat luokat ovat tiedossa.

4.2 Ennustava ja selittävä mallinnus

Luokittelu on tutkimusmenetelmänä ennustavaa mallintamista (*predictive modeling*): mallin avulla pyritään siis ennustamaan oikea luokka Y syötteelle X. Tämä eroaa monesti tilastotieteessä käytettävästä selittävästä mallintamisesta (*explanatory modeling*), jossa perusoletuksena on, että tekijät X johtuvat muuttujista Y, ja mallintamalla pyritään selittämään tätä kausaalista yhteyttä. Ennustavan ja selittävän mallinnuksen etujen välinen debatti on paitsi tilastotieteellinen, myös tieteenfilosofinen (kts. esim. Forster & Sober, 1994).

Breiman (2001) on vertaillut tiedon mallinnusta ja algoritmista mallinnusta (*data modeling* ja *algorithmic modeling*). Breiman esittää, että ns. perinteisessä tiedon mallinnuksessa, kuten logistisessa regressioanalyysissä, ulotteisuutta usein pienennetään poistamalla muuttujia, sillä

laaja ulotteisuus johtaa herkästi kohinaan ja tekee mallinnuksesta haastavaa. Modernit teknologiat kuitenkin poistavat tämän ongelman ja esimerkiksi tukivektorikonetta käyttävässä ennustavassa mallinnuksessa voi olla tuhansia ulottuvuuksia ja piirteitä, joiden avulla ilmiötä selvitetään. Tämä tekee tukivektorikoneesta houkuttelevan juuri kielentutkimuksen näkökulmasta: teksti ja kieli ovat luonnostaan erittäin moniulotteisia, kun esimerkiksi yksittäiset sanat nähdään eri ulottuvuuksina. Tiedon mallinnuksen keinojen, kuten logistisen regression, etuina on selitettävyys ja tulkittavuus, mutta algoritmisen mallinnuksen avulla voidaan päästä korkeampaan tarkkuuteen. Nykyään koneoppimismenetelmätkään eivät ole enää täysin tulkittavuuden ulottumattomissa, ja esimerkiksi tukivektorikoneiden tekemien luokittelujen selitettävyyteen on esitetty useita keinoja, joista tarkemmin luvussa 4.4.

Ennustavan ja selittävän mallintamisen eroista on kirjoittanut esimerkiksi Shmueli (2010). Shmuelin mukaan ennustavan mallintamisen etuna on se, että sen avulla voidaan saada näkyviin X:n ja Y:n välisiä, erittäin monimutkaisiakin yhteyksiä vaikka tietoa kausaalisista syistä ilmiön taustalla ei olisi mahdollista selvittää.

Tässä tutkimuksessa ei keskitytä siihen, miksi eroja sukupuolten välillä on politiikan uutisoinnissa, vaan siihen, onko eroja ja jos on, millaisia ne ovat. Ennustava mallinnus antaa tähän vastauksia ilman, että aineistoa tarvitsee yksinkertaistaa selittävän mallinnuksen yleensä vaatimalla tavalla. Modernit koneoppimisen menetelmät ovat tehneet ennustavasta, algoritmisesta mallinnuksesta robustia.

4.3 Luokittelu tekstianalyysin metodina

Tekstin luokittelua on käytetty eri tavoin tekstin piirteiden analysointiin. Luokittelija ensin koulutetaan luokittelemaan tekstejä annettuihin luokkiin, ja mikäli luokittelija onnistuu tehtävässä tarpeeksi hyvin eli yli perustason, voidaan luokittelijan luokittelussa käyttämiä piirteitä analysoida eri luokat erottavina piirteinä. Tällaista piirreanalyysiä on tehty tukivektorikoneita käyttäen myös muuhun dataan kuin tekstiin. Tukivektorikoneet on useissa tutkimuksissa todettu erittäin tehokkaaksi tavaksi piirteiden löytämiseen ja erotteluun (esim. Joachims, 1998; Guyon ym., 2002; Forman, 2003).

Diermeier ym. käyttivät tätä metodia tutkiakseen Yhdysvaltain kongressissa pidettyjä puheita ja erotellakseen piirteitä, joiden perusteella luokittelija erotteli puheen konservatiiviseksi tai liberaaliksi (Diermeier ym., 2011). Tutkimuksessa esitettiin metodi, jossa tukivektorikoneen

generoimat piirteet laitettiin järjestykseen niiden kertoimien (*coefficient*) mukaan. Kaksiluokkaisessa luokittelussa toinen luokka on laskennallisesti positiivinen ja toinen negatiivinen, joten esimerkiksi Diermeier ym. tutkimuksessa mitä suurempi positiivinen kerroin, sitä voimakkaammin piirre viittasi konservatiiviseen luokkaan ja vastaavasti negatiivinen kerroin viittasi liberaaliin luokkaan. Näitä piirteitä analysoimalla saadaan esiin kielellisiä ominaisuuksia, jotka ovat sidoksissa luokkiin. Tätä samaa menetelmää käytetään tässä tutkimuksessa.

Edellä lyhyesti esitellyssä tutkimuksessa luokittelu liittyi tekstin tuottajaan eli puheen pitäjän ideologiaan. Tukivektorikoneita on käytetty vastaavalla tavalla luokittelemaan tekstejä kirjoittajan sukupuolen perusteella (esim. Argamon ym., 2009; Montero ym., 2014).

Samaa menetelmää on käytetty myös tutkimuksissa, jotka eivät keskity tekstin tuottajaan vaan tekstin muihin piirteisiin. Tukivektorikoneluokittelijaa on käytetty esimerkiksi tekstilajien tunnistukseen. Sekä kieliopillisia että sanastollisia piirteitä käyttänyt tukivektorikone pystyi luokittelemaan Laippala ym. tutkimuksessa erittäin laajan CORE-korpuksen englanninkielisiä internet-tekstejä 26 eri tekstilajiin (2021). Tutkimus osoitti, että sanastolliset piirteet olivat tärkeitä erottelevia piirteitä osalle tekstilajeista. Jo aiemmin esitelty Leavyn (2019) tutkimus käytti tukivektorikoneluokittelijaa selvittääkseen sukupuolten välisiä eroja politiikan uutisoinnissa ja käytti kvalitatiivista analyysiä erottelevien piirteiden tutkimisessa. Menetelmä on tuttu edellä esitetystä Diermeier ym. tutkimuksesta (2011).

4.4 Luokittelijan arvioinnista

Luokittelijan arviointiin on useampia erilaisia keinoja, mutta yleisimmin käytetyt metriikat ovat täsmällisyys (*accuracy*), tarkkuus (*precision*), herkkyys (*recall*), ja F1-mitta (*F1-score*). Näiden lisäksi esitetään yleensä luokittelijan sekaannusmatriisi (*confusion matrix*).

Binäärisellä luokittelijalla on käytössään kaksi luokkaa. Luokista käytetään termejä positiivinen ja negatiivinen: luokittelija siis arvioi esimerkiksi, onko teksti runo vai ei. Luokittelija voi tehdä neljä erilaista päätelmää: oikea positiivinen (luokittelija ennustaa runon runoksi), väärä positiivinen (teksti ei ole runo, mutta luokittelija ennustaa sen olevan), oikea negatiivinen (teksti ei ole runo eikä luokittelija sitä sellaiseksi ennusta) ja väärä negatiivinen (teksti on runo, mutta luokittelija ennustaa, ettei se ole).

Sekaannusmatriisi on keino luokittelijan täsmällisyyden arviointiin. Siinä esitetään oikeiden positiivisten ja negatiivisten sekä väärien positiivisten ja negatiivisten määrä.

Taulukko 1. Sekaannusmatriisi binäärisessä luokittelussa

Oikea luokka / Ennustettu luokka	Positiivinen	Negatiivinen
Positiivinen	Oikea positiivinen	Väärä positiivinen
Negatiivinen	Väärä negatiivinen	Oikea negatiivinen

Täsmällisyys kuvaa sitä, kuinka suuri osa luokittelijan ennustamista luokista meni oikein. Se on siis oikein ennustettujen ja kaikkien ennustusten suhdeluku.

$$\frac{oikeat\ positiiviset + oikeat\ negatiiviset}{oikeat\ positiiviset + oikeat\ negatiiviset + väärät\ positiiviset + väärät\ negatiiviset}$$

Tarkkuus on oikein ennustettujen positiivisten suhde kaikkiin positiiviseksi ennustettujen määrään. Se kuvaisi runoesimerkkiä jatkaen siis sitä, mikä osa runoiksi ennustetuista runoista todella oli runoja.

$$\frac{oikeat\ positiiviset}{oikeat\ positiiviset + väärät\ positiiviset}$$

Herkkyys puolestaan vertaa oikeiden positiivisten määrää siihen lukuun, kuinka monta positiivista koko aineistossa oli: kuinka monta runoa luokittelija siis tunnisti runoksi kaikista aineiston runoista.

$$\frac{oikeat\ positiiviset}{oikeat\ positiiviset + väärät\ negatiiviset}$$

Näiden lisäksi käytetään vielä F1-mittaa, joka on harmoninen keskiarvo kaikista edellä esitetyistä mitoista.

$$F1 = 2 \times \frac{tarkkuus \times herkkyys}{tarkkuus + herkkyys}$$

Kun luokiteltavana on useita luokkia, periaate pysyy samana. Kaavoissa kolmea eri luokkaa hahmotetaan kirjaimilla A, B ja C.

Tarkkuus lasketaan moniluokkaisessa luokittelussa käytännössä samalla tavalla kuin binäärisessäkin luokittelussa. Kyseessä on oikein luokiteltujen suhde kaikkiin luokittelijan kyseiseen luokkaan luokiteltuihin. Tarkkuus lasketaan jokaiselle luokalle erikseen.

$$\frac{oikeat\ A{:}\ t}{oikeat\ A{:}\ t + väärät\ A{:}\ t}$$

Herkkyyttä laskiessa verrataan oikein luokiteltujen suhdetta kaikkiin luokkaan kuuluviin samoin kuin binäärisessä luokittelussa. Myös herkkyys lasketaan jokaiselle luokalle erikseen.

$$\frac{oikeat\ A{:}\ t}{oikeat\ A{:}\ t + B{:}\ ksi\ tai\ C{:}\ ksi\ luokitellut, jotka\ oikeasti\ A}$$

Kun arvioidaan sitä, mitkä täsmällisyys- tai tarkkuusarvot ovat tarpeeksi hyviä, voidaan käyttää vertailussa jotakin tiettyä perustasoa (*baseline*). Yksinkertaisimmillaan perustasoksi voidaan binäärisessä luokittelussa valita testidatan luokkien jakauma: jos aineisto jakautuu tasan molempiin luokkiin, olisi perustaso tällöin 50 % (tai 0.5). Mallin tulisi pystyä tätä parempaan täsmällisyyteen. Jos aineisto ei ole jakautunut tasan, vaan toista luokkaa on esimerkiksi 60 %, perustasona voisi olla 0.6 – se olisi täsmällisyys, jonka saisi malli, joka aina ennustaisi enemmistöluokan syötteen piirteistä huolimatta.

5 Aineisto ja sen käsittely

Aineistona on RSS-syötteestä poimittuja Yle Uutiset -sivustolla julkaistuja uutisia sekä samaan tapaan poimittuja Helsingin Sanomien internet-sivuilla julkaistuja uutisia. Kaikki aineiston uutiset ovat vapaasti saatavilla, eli aineisto ei sisällä ns. maksumuurin takana olevia uutisia.

Oxfordin yliopiston Reuters-instituutin vuoden 2021 Digital News Report -tutkimuksen Suomen maaraportin mukaan 44 prosenttia suomalaisista kertoo seuraavansa Ylen nettiuutisia (ml. Yle Areenassa olevat uutiset) vähintään viikoittain ja 30 prosenttia vähintään kolme kertaa viikossa (Reunanen ym., 2021). Ylen edelle pääsee raportissa vain kaksi verkkomediaa, Iltalehti ja Ilta-Sanomat. Helsingin Sanomien nettiuutisia saman raportin mukaan luki viikossa 30 prosenttia suomalaisista ja vähintään kolmesti viikossa 21 prosenttia. Sekä Ylen että Helsingin Sanomien verkkouutiset ovat siis erittäin luettuja, mikä tukee niiden valintaa tutkimusaineistoksi. Näiden medioiden uutisin myös luotetaan Digital News Reportin mukaan eniten tutkituista suomalaisista medioista: Ylen uutisiin kertoi luottavansa 85 prosenttia ja Helsingin Sanomien 81 prosenttia (Reunanen ym., 2021).

5.1 Eduskunta ja poliittinen toimintaympäristö aineiston aikavälillä

Uutiset ovat aikaväliltä 31.12.2019–31.12.2021. Edelliset eduskuntavaalit järjestettiin vuonna 2019 ja uusi eduskunta aloitti työnsä 23.4.2019. Aineiston aikajänne on siis sellainen, ettei sen aikana ole järjestetty vaaleja ja eduskunnan kokoonpano pysyi koko ajan samana. Ainoa kokoonpanoon liittyvä muutos tällä aikavälillä oli Ano Turtiaisen erottaminen Perussuomalaisista, mutta Turtiainen pysyi kansanedustajana ensin sitoutumattomana ja 5.6.2021 eteenpäin Valta kuuluu kansalle -ryhmässä. Lista käytetyistä 200 kansanedustajan nimestä löytyy liitteistä. Lista on haettu Eduskunnan sivuilta[4] 8.4.2022.

Vuoden 2019 vaaleissa eniten paikkoja (40) sai Suomen Sosialidemokraattinen Puolue (SDP), toiseksi eniten Perussuomalaiset (PS, 39 paikkaa) ja kolmanneksi eniten Kansallinen Kokoomus (Kok, 38 paikkaa) (Yle, ei pvm.). Vaaleissa valittiin eduskuntaan 94 nais- ja 106 miesehdokasta. Luvut ovat poikkeukselliset, sillä naisia valittiin vuoden 2019 vaaleissa

[4] https://www.eduskunta.fi/FI/kansanedustajat/Sivut/Kansanedustajat-aakkosjarjestyksessa.aspx

eduskuntaan enemmän kuin koskaan aiemmin (Konttinen, 2019). Naisehdokkaista eniten ääniä sai Vasemmistoliiton Li Andersson ja miehistä Perussuomalaisten Jussi Halla-aho (Yle, ei pvm.). Aineiston aikavälillä edustajista 108 oli miehiä ja 92 naisia; vaalien jälkeen muutoksia kokoonpanossa siis tapahtui, mutta ei aineiston aikavälillä.

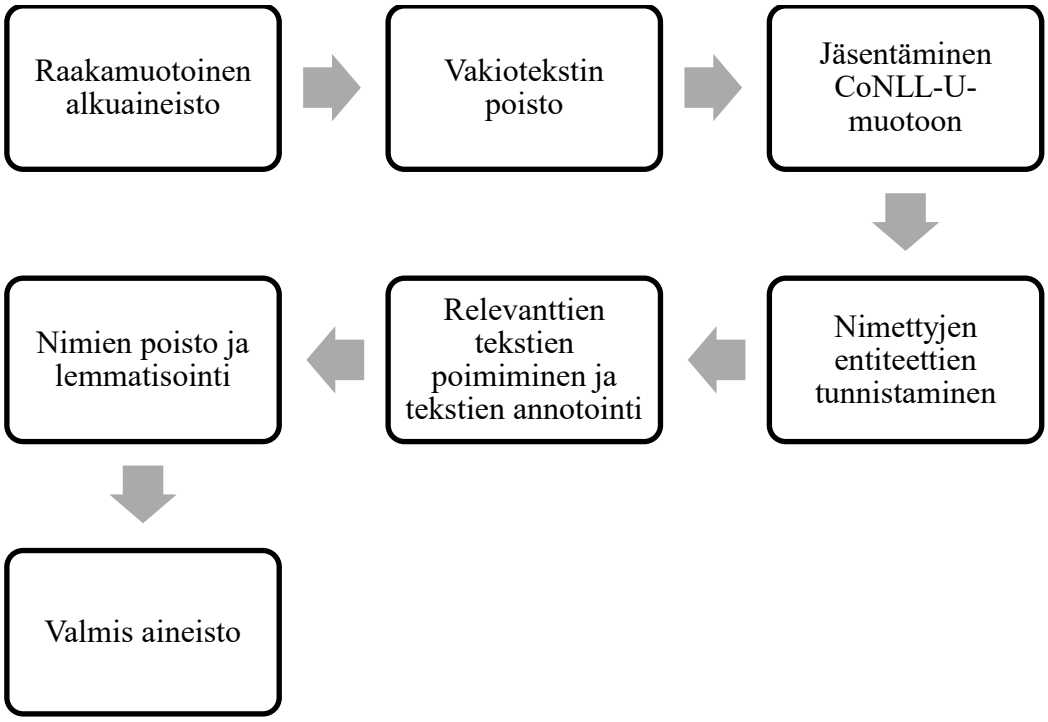
Hallituksen muodostivat Suomen Sosialidemokraattinen Puolue, Suomen Keskusta, Vihreä Liitto, Vasemmistoliitto ja Suomen ruotsalainen kansanpuolue. Vaalien jälkeen 6. kesäkuuta 2019 pääministeriksi valittiin SDP:n Antti Rinne, joka erosi tehtävästään saman vuoden joulukuussa. Hänen jälkeensä pääministeriksi nousi SDP:n Sanna Marin, ja Marinin hallitus aloitti työnsä 10.12.2019 (Valtioneuvosto, ei pvm.).

Merkittävimpiä ajanjaksoa 31.12.2019–31.12.2021 leimaavia tapahtumia, jotka heijastuivat juuri politiikan uutisointiin, olivat erilaiset koronapandemiaa koskevat päätökset. Tammikuun lopussa 2020 Suomessa havaittiin maan ensimmäinen koronavirustartunta (Ruokangas ym., 2020) ja nopeasti tilanne eskaloitui niin, että maaliskuun 13. päivänä 2020 hallitus otti valmiuslain käyttöön ja myöhemmin saman kuun aikana Uudenmaan raja suljettiin muulta kuin välttämättömältä liikenteeltä (Valtioneuvosto, 2020). Pandemian hoito, erilaiset rajoitukset ja muut siihen liittyvät poliittiset päätökset olivat usein uutisissa koko edellä mainitun ajanjakson ajan. Pandemiaan liittyvissä aiheissa näkyvyyttä sai etenkin Marinin hallitus, jota johti pandemian alkuaikoina viisi naisministeriä: pääministeri Sanna Marin, valtiovarainministeri Katri Kulmuni, sisäministeri Maria Ohisalo, opetusministeri Li Andersson ja oikeusministeri Anna-Maja Henriksson. Pandemia-asioissa paljon esillä olivat myös perhe- ja peruspalveluministeri Krista Kiuru sekä sosiaali- ja terveysministeri Aino-Kaisa Pekonen.

5.2 Aineiston esiprosessointi

Yle Uutisten uutisia aineistossa oli 53 167 uutista ja Helsingin Sanomien aineistossa 96 513 uutista. Kaikki uutiset eivät kuitenkaan käsittele politiikkaa, joten aineistoa on täytynyt rajata ja käsitellä, jotta lopulliseen analyysiin päätyy ainoastaan relevantteja tekstejä. Jotta näitä uutisartikkeleita voidaan käyttää tässä tutkimuksessa, on aineistoa esiprosessoitava ennen varsinaista tutkimusta ja analyysiä. Koneoppimismalli tarvitsee koulutusdatakseen sellaista tekstiä, jossa käsitellään poliitikkoja. Koska kyseessä on luokittelija, tarvitsee se myös tekstien oikeat luokat: tässä tapauksessa luokat nainen ja mies. Aineistosta on siis poimittava tutkimuksen kannalta relevantit tekstit ja käsiteltävä niitä niin, että luokittelija tekee

luokittelun muiden kuin kansanedustajan nimen perusteella, jotta koneoppimismalli ei yksinkertaisesti oppisi luokittelemaan Sanna-nimen sisältävät tekstit luokkaan nainen ja Pekka-nimen sisältävät luokkaan mies.



Kuvio 1. Vuokaavio esiprosessoinnista

Yllä oleva vuokaavio kuvaa esiprosessoinnin eri vaiheet, jotka käyn seuraavaksi tarkemmin läpi.

Uutiset ovat HTML-muodossa, ja sellaisenaan niiden tutkiminen koneoppimismenetelmin ei ole järkevää: ne sisältävät paljon vakiotekstiä (engl. *boilerplate*), eli linkkejä, navigointivalikkoja, HTML-koodia ja niin edelleen. Vakiotekstin poistamiseen on käytetty Trafilatura-nimistä ohjelmistoa (Barbaresi, 2021). Tekstien yhteyteen säilytettiin niiden metadata, eli tiedot esimerkiksi otsikosta, julkaisupäivästä ja tekstin kirjoittaneesta toimittajasta. Analyysissä on käytetty ainoastaan leipätekstiä, eli otsikko ja mahdolliset kuvatekstit eivät ole mukana.

Kuva 1. Esimerkki siitä, miltä HTML-tiedosto näyttää ennen Trafilatura-ohjelmalla puhdistamista.

Itä-Suomen hovioikeus on koventanut Heinolan ja Hämeenlinnan amfetamiinikaup
pajutun päätekijän tuomiota. Hovioikeus korotti vankeusrangaistuksen pituude
ksi kymmenen vuotta, kun Päijät-Hämeen käräjäoikeus oli tuominnut miehen kah
deksan vuoden ja kymmenen kuukauden vankeuteen.
Hovioikeuden mukaan tuomitulla oli keskeinen rooli Cannonball-moottoripyöräj
engissä Heinolan ja Hämeenlinnan kerhoissa tehtyjen rikosten suunnittelussa.
Rikokset tapahtuivat kesästä 2016 kesään 2017 ulottuvalla ajanjaksolla Hämee
nlinnan, Heinolan ja Janakkalan seuduilla.
Oikeuden mukaan Cannonballin Heinolan-osastoon kuuluva jäsen luovutti 12 kil
oa amfetamiinia Cannonballin Hämeenlinnan-osaston kokelasjäsenelle toukokuus
sa 2017. Sen jälkeen Cannonballin johtoon kuuluvat kaksikko järjesti ja joht
i huumeiden myymistä ja levittämistä edelleen.

Kuva 2. Sama tiedosto Trafilatura-puhdistuksen jälkeen.

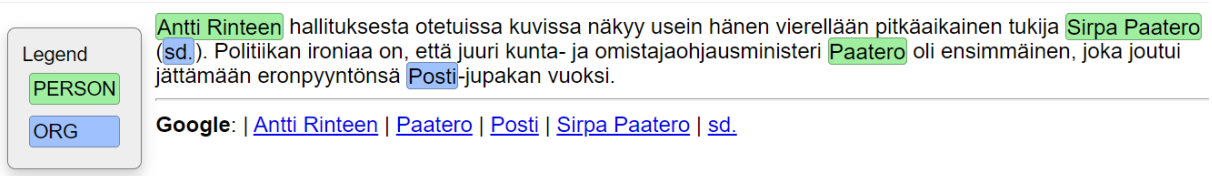
Puhdistamisen jälkeen aineisto jäsennettiin CoNLL-U-muotoon käyttäen Turku neural parser pipeline -jäsennintä (Kanerva ym., 2018). CoNLL-U-muodossa joka sanasta on esillä runsas määrä tietoa, esimerkiksi sanan perusmuoto eli lemma, sanan sijamuoto ja niin edelleen. CoNLL-U-muodon hyötynä oli tämän tutkimuksen kannalta se, että aineiston kaikista virkkeistä saatiin sekä alkuperäiset että lemmatisoidut versiot.

Taulukko: esimerkki siitä, miltä lause näyttää CoNLL-U-muotoon jäsennettynä

# text = Marinin mukaan näin oli ehdottomasti syytä toimia.									
1	Marinin	Marin	PROPN	_	Case=Gen|Number=Sing	6	obl	_	_
2	mukaan	mukaan	ADP	_	AdpType=Post	1	case	_	_
3	näin	näin	ADV	_	_	7	advmod	_	_
4	oli	olla	AUX	_	Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin|Voice=Act	6	cop	_	_
5	ehdottomasti	ehdottomasti	ADV	_	Derivation=Sti	6	advmod	_	_

6	syytä	syy	NOUN	_	Case=Par|Number=Sing	0	root	_	_
7	toimia	toimia	VERB	_	InfForm=1|Number=Sing|VerbForm=Inf|Voice=Act	6	xcomp:ds	_	SpaceAfter=No
8	.	.	PUNCT	_	_	6	punct	_	SpacesAfter=\n

Aineistosta poimittiin analyysiin vain ne artikkelit, jotka olivat tutkimuksen kannalta oleellisia, eli ne, joissa mainitaan nimeltä yksi tai useampi kansanedustaja. Tämän mahdollistamiseksi aineistosta on pitänyt pystyä tunnistamaan juuri ne tekstit, joissa mainitaan kansanedustaja. Koska suomen kielessä nimet voivat esiintyä taivutettuina, pelkän nimilistan läpikäynti ei ole mahdollista. Vaikka etsittäisiin tekstien joukosta niitä, joissa mainitaan esimerkiksi nimi ”Antti Rinne”, artikkelit, joissa nimi esiintyy ainoastaan muodossa ”Antti Rinteen”, eivät tulisi valituksi aineistoon. Tämä on ratkaistu käyttämällä suomen kieltä varten tehtyä NER-ohjelmaa (Luoma ym., 2020). NER eli *Named Entity Recognition* tarkoittaa nimettyjen entiteettien tai kokonaisuuksien tunnistamista eli automatisoitua ihmisten nimien, päivämäärien, organisaatioiden, paikkojen ja muiden vastaavien tunnistamista tekstistä.



Kuva 3. Finnish NER -tunnistimen (Luoma ym., 2020) demoversio havainnollistaa, miten tunnistin merkitsee eri nimetyt entiteetit tekstiin.

Edellä kuvatun prosessin jälkeen tunnistimen PERSON-tunnisteen saaneiden nimien perusteella aineistoon voitiin valita ainoastaan relevantit artikkelit. Koska CoNLL-U-muodon ansiosta aineiston kaikista sanoista oli tallessa niiden lemmatisoidut muodot, PERSON-tunnisteen saaneista nimistä voitiin hakea CoNLL-U-muodosta niiden perusmuotoiset versiot ja verrata perusmuodossa olevaa nimeä listaan kansanedustajien nimistä. Näin tekemällä siis saadaan aiemman esimerkin ”Antti Rinteen” -sanaparin sisältämä teksti saadaan mukaan aineistoon, sillä CoNLL-U-muodossa sama sanapari on ”Antti Rinne”.

Kansanedustajien nimet on saatu Suomen eduskunnan sivuilta, josta löytyy luettelot sekä nykyisistä että kaikista entisistä kansanedustajista. On mahdollista, että aineistoon on päätynyt tekstejä, joissa ei mainita kansanedustajaa vaan haastateltavana on sattumalta kansanedustajan

kanssa saman niminen henkilö. Kansanedustajan päätyminen Ylen Uutisiin voidaan kuitenkin arvioida niin paljon mahdollisia kaimojansa todennäköisemmäksi, etten usko tämän vääristävän tuloksia.

Yhdessä tämän vaiheen kanssa aineisto annotoitiin. Jotta koneoppimismalli voidaan kouluttaa, se tarvitsee valmiiksi annotoidun opetusdatan, eli syötteillä täytyy olla valmiina oikea luokka. Tätä merkintää kutsutaan tekstin leimaksi. Tässä tutkimuksessa leima on ollut joko ”M” tai ”F”, viitaten englanninkielisiin sanoihin *male* ja *female*. Annotointi tehtiin automatisoidusti nimilistan perusteella. Nimilista leimoineen on tämän tutkielman liitteenä.

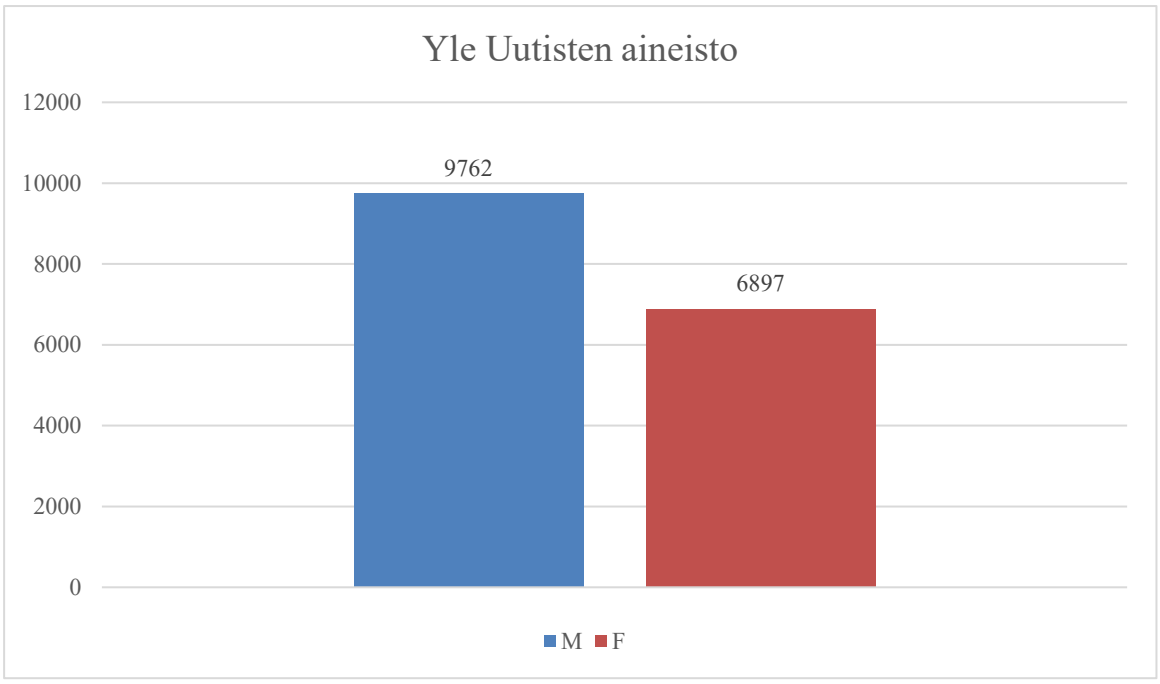
Seuraavaksi teksteistä piti poistaa kansanedustajien nimet. Koneoppimismallia testattiin niin, että nimiä ei ollut poistettu tekstistä, ja eri kansanedustajien nimet nousivat silloin erittäin merkittäviksi tekijöiksi luokittelussa. Malli siis oppi, että teksti ”Sanna Marin” nostaa nainen-luokan todennäköisyyttä merkittävästi ja vastaavasti esimerkiksi ”Pekka Haavisto” teki saman mies-luokalle. Kaikki teksteissä esiintyvät kansanedustajien nimet on siten korvattu yksinkertaisesti ”Nimi”-tekstillä, jotta muut mahdolliset tekstin piirteet pääsisivät paremmin esille nimien sijaan. Tekstien sanat on myös lemmatisoitu, jotta eri taivutusmuodot samasta sanasta eivät olisi eri piirteitä, vaan malli tulkitsisi ne samaksi piirteeksi. Poimin lopulliseen aineistoon uutisista sellaiset virkkeet, joissa mainitaan kansanedustaja joko koko nimellä tai sukunimellä. Käyttämällä datana tällaisia virkkeitä, saadaan eriteltyä piirteitä, jotka ovat erittäin läheisesti yhteydessä poliitikkoihin (Leavy, 2019).

Seuraavassa taulukossa on esimerkki siitä, miltä aineisto näyttää nimien poistamisen ja lemmatisoinnin jälkeen.

Leima	Teksti
M	SDP Nimi Nimi arvioida torstai illansuu hallitusneuvottelu lähteä kulkea suunnitella mukaisesti .
F	– Suomi erityisolo ja suomalainen maatalous puolustaminen rahoituskehysneuvottelu , vastata Nimi Nimi .
M	keskusta lähettää hallituskysely puoluevaltuusto , puoluehallitus ja eduskuntaryhmä 177 jäsen tiistai-ilta kello 18.43. SDP puheenjohtaja Nimi Nimi olla informoida keskusta olettaa voimasuhde jo ennen virallinen kysely .
M	– lobbarikeskustelu olla mennä yli äyräs , Nimi summata päivä tapahtuma .
F	kun Nimi kysyä , mikä hän tulla iso , tyttö heläyttää : kuuluisa – ja naurata päälle .
F	helsinkiläinen Nimi , 34 , valita eduskunta Helsinki runsas 11 000 ääni .
M	vihre puheenjohtaja Nimi Nimi uskoa , että hanke olla saada EUraha .
M	toinen huolenaihe Nimi pitää väittää lobbari suuri määrä neuvottelu .
F	toisaalta juttu aihe olla Nimi mukaan myös " vähän arka " , mikä se saattaa olla osasyy .
F	kansanedustaja Nimi Nimi olla hyvin todennäköisesti vihre seuraava puheenjohtaja .

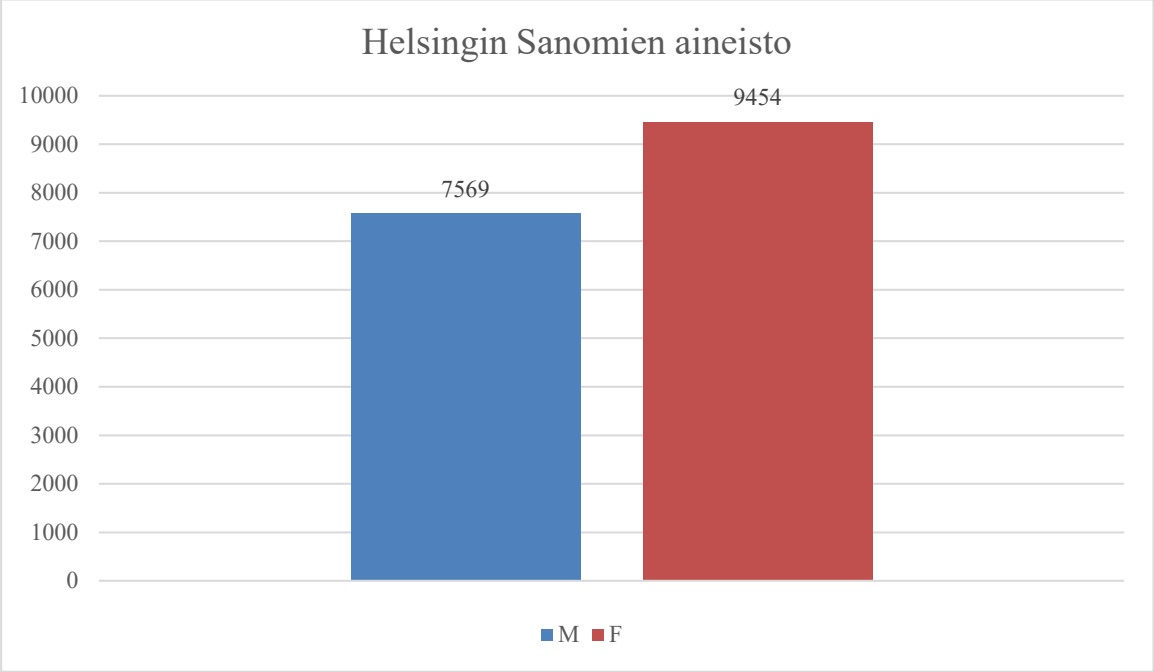
Välimerkit on erotettu välilyönnein, jotta malli ei tulkitsisi välimerkin kuuluvan sitä edeltävään sanaan.

5.3 Aineistot lukuina



Kuvio 2. Ylen aineiston sukupuolijakauma

Yle Uutisten osalta 53 167 uutisartikkelia sisältävästä aineistosta lopulliseen aineistoon kansanedustajan nimeltä mainitsevia virkkeitä aineistoon tuli yhteensä 16 659, joista 9 762 edusti luokkaa M eli mies ja 6 897 luokkaa F eli nainen. Myöhemmin analyysissä luokat esitetään muodossa 0 = nainen ja 1 = mies.



Kuvio 3. Helsingin Sanomien aineiston sukupuolijakauma

Helsingin Sanomien aineiston osalta 96 513 uutisartikkelin esiprosessoinnin jälkeen virkkeitä päätyi aineistoon yhteensä 17 023, joista 9 454 edusti luokkaa F eli nainen ja 7 569 luokkaa M eli mies. Toisin kuin Ylen aineistoissa, Helsingin Sanomien aineistossa F-luokkaa on siis enemmän.

Tutkimuksessa käytetyn kansanedustajalistan nimistä 92 kuului luokkaan nainen ja 108 luokkaan mies. Prosentuaalisesti jakauma on siis 46 prosenttia naisia ja 54 prosenttia miehiä. Yle Uutisten aineiston virkkeiden jakauma on 41 prosenttia naisia ja 59 prosenttia miehiä, Helsingin Sanomien osalta 56 prosenttia naisia ja 44 prosenttia miehiä. Yle Uutisten jakauma on siis lähempänä kansanedustajien todellista sukupuolijakoa.

5.4 Koneoppimismallin koulutus

Jatkoprosessoinnissa kaikki lauseiden teksti muutettiin pienillä kirjaimilla kirjoitetuksi, jotta koneoppimismalli ei pitäisi esimerkiksi sanoja ”Politiikka” ja ”politiikka” eri piirteinä. Sen jälkeen lauseet tokenisoitiin, eli ne muutettiin pitkistä yhtenäistä teksteistä listoiksi tokeneita eli käytännössä sanoja. Näistä listoista poistettiin vielä hukkasanat (*stop words*). Suomenkielisten hukkasanojen listassa on esimerkiksi persoonapronominit taivutusmuotoineen, olla-verbin eri muodot, konjunktioita sekä muita pieniä, yleisiä sanoja kuten yli, nyt ja niin. Myös numerot on poistettu, kirjoitetut numeraalit on jätetty.

Tämän jälkeen vektoroin aineiston käyttämällä TF-IDF-vektorointia. Lyhenne TF-IDF tulee sanoista *Term Frequency – Inverse Document Frequency*. Tässä vaiheessa sanalistat muutettiin siis vektoreiksi eli numeeriseen muotoon. Vektorointi tehtiin niin, että piirteitä saattoi tulla maksimissaan 5000. Tämä tarkoittaa sitä, että vektoroijan muodostaman sanaston suuruus on 5000, eikä se siten välttämättä ota huomioon jokaista erillistä sanaa joka aineistossa esiintyi. Valinnat siitä, mitkä sanat päätyvät 5000 joukkoon, vektoroija tekee asettamalla piirteet järjestykseen *term frequencyn* eli termien tiheyden perusteella. Tämän ansiosta sellaiset termit eli piirteet, jotka ovat erittäin harvinaisia ja esiintyvät esimerkiksi vain kerran, eivät ole mukana piirteiden tarkastelussa. Jos tällaiset hyvin harvinaiset piirteet olisivat mukana, nostaisi sanan esiintyminen vain kerran tai kaksi ja aina yhteydessä joko mies- tai nainen-luokkaan luokitelluissa virkkeissä sen niin sanotusti kärkeen piirteitä tarkastellessa luoden kuitenkin väärää kuvaa, jos sana ei ole uutisoinnissa mitenkään yleinen. Näin ollen kaikki myöhemmässä tarkastellussa ovat piirteet ovat siis esiintyneet aineistossa useaan kertaan.

Koneoppimismallina aineistoon sovitettiin SVM:ää käyttäen Scikit learn -kirjaston LinearSVC-moduulia[5]. Tärkeimpiä hyperparametreja tässä moduulissa on C, joka optimoi tukivektorikoneen tekemää erottelua.

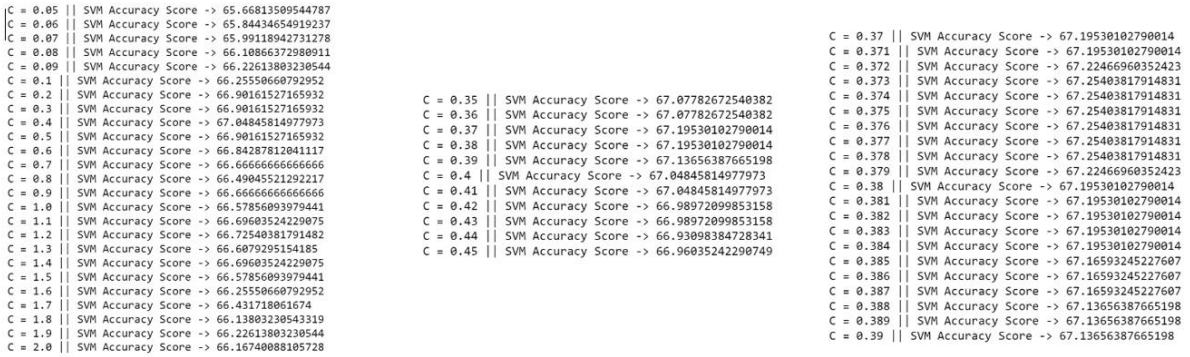
```
C = 0.05 || SVM Accuracy Score -> 65.78631452581033
C = 0.06 || SVM Accuracy Score -> 65.93637454981993
C = 0.07 || SVM Accuracy Score -> 66.20648259303722
C = 0.08 || SVM Accuracy Score -> 66.71668667466987
C = 0.09 || SVM Accuracy Score -> 66.83673469387756
C = 0.1 || SVM Accuracy Score -> 66.92677070828331
C = 0.2 || SVM Accuracy Score -> 66.59663865546219
C = 0.3 || SVM Accuracy Score -> 66.8967587034814
C = 0.4 || SVM Accuracy Score -> 66.59663865546219
C = 0.5 || SVM Accuracy Score -> 66.53661464585834
C = 0.6 || SVM Accuracy Score -> 66.6266506602641
C = 0.7 || SVM Accuracy Score -> 66.47659063625451
C = 0.8 || SVM Accuracy Score -> 66.0564225690276
C = 0.9 || SVM Accuracy Score -> 66.17647058823529
C = 1.0 || SVM Accuracy Score -> 66.02641056422569
C = 1.1 || SVM Accuracy Score -> 65.96638655462185
C = 1.2 || SVM Accuracy Score -> 65.90636245018
C = 1.3 || SVM Accuracy Score -> 65.75630252100841
C = 1.4 || SVM Accuracy Score -> 65.66626650660264
C = 1.5 || SVM Accuracy Score -> 65.6062424969988
C = 1.6 || SVM Accuracy Score -> 65.69627851140456
C = 1.7 || SVM Accuracy Score -> 65.57623049219687
C = 1.8 || SVM Accuracy Score -> 65.69627851140456
C = 1.9 || SVM Accuracy Score -> 65.63625450180072
C = 2.0 || SVM Accuracy Score -> 65.54621848739495
```

```
C = 0.105 || SVM Accuracy Score -> 66.92677070828331
C = 0.106 || SVM Accuracy Score -> 66.8967587034814
C = 0.107 || SVM Accuracy Score -> 67.01680672268907
C = 0.108 || SVM Accuracy Score -> 67.046818727491
C = 0.109 || SVM Accuracy Score -> 66.95678271308523
C = 0.11 || SVM Accuracy Score -> 66.95678271308523
C = 0.111 || SVM Accuracy Score -> 66.92677070828331
C = 0.112 || SVM Accuracy Score -> 66.8967587034814
C = 0.113 || SVM Accuracy Score -> 66.8967587034814
C = 0.114 || SVM Accuracy Score -> 66.8967587034814
C = 0.115 || SVM Accuracy Score -> 66.8967587034814
```

```
C = 0.1068 || SVM Accuracy Score -> 67.01680672268907
C = 0.1069 || SVM Accuracy Score -> 67.01680672268907
C = 0.107 || SVM Accuracy Score -> 67.01680672268907
C = 0.1071 || SVM Accuracy Score -> 67.046818727491
C = 0.1072 || SVM Accuracy Score -> 67.01680672268907
C = 0.1073 || SVM Accuracy Score -> 67.01680672268907
C = 0.1074 || SVM Accuracy Score -> 67.01680672268907
C = 0.1075 || SVM Accuracy Score -> 67.046818727491
C = 0.1076 || SVM Accuracy Score -> 67.046818727491
C = 0.1077 || SVM Accuracy Score -> 67.046818727491
C = 0.1078 || SVM Accuracy Score -> 67.046818727491
C = 0.1079 || SVM Accuracy Score -> 67.046818727491
C = 0.108 || SVM Accuracy Score -> 67.0468187274 91
C = 0.1081 || SVM Accuracy Score -> 67.046818727491
C = 0.1082 || SVM Accuracy Score -> 66.98679471788715
C = 0.1083 || SVM Accuracy Score -> 66.98679471788715
C = 0.1084 || SVM Accuracy Score -> 66.98679471788715
C = 0.1085 || SVM Accuracy Score -> 66.98679471788715
C = 0.1086 || SVM Accuracy Score -> 66.98679471788715
C = 0.1087 || SVM Accuracy Score -> 66.98679471788715
C = 0.1088 || SVM Accuracy Score -> 66.95678271308523
C = 0.1089 || SVM Accuracy Score -> 66.95678271308523
```

Kuva 1. Yle Uutisten aineiston optimaalin C-arvon etsintä

[5] https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html



Kuva 2. Helsingin Sanomien aineiston optimaalin C-arvon etsintä

Yllä olevista kuvista näkyy, miten tukivektorikoneen täsmällisyys (SVM Accuracy Score) muuttuu eri C:n arvoilla. Haarukoinnin lopputuloksena Yle Uutisten aineistolla tarkimman tuloksen tuovat C:n arvot 0.1075–0.1081 ja Helsingin Sanomien aineistossa 0.373–0.378. Päädyin käyttämään Ylen aineiston osalta C:n arvoa 0.1078 ja Helsingin Sanomien aineistolla 0.375.

Lopullista mallin koulutusta varten aineisto on jaettu koulutus- ja testausdataan niin, että 80 prosenttia aineistosta käytettiin koulutukseen ja 20 prosenttia testaukseen. Mallin koulutuksen jälkeen on mahdollista tarkastella tutkimuksen varsinaisia tuloksia, eli analysoida niitä piirteitä, joiden perusteella malli teki luokittelun.

6 Tulokset ja analyysi

Tässä luvussa esittelen tutkimuksen tuloksia käyttäen aiemmin luvussa 4.3 esiteltyjä tunnuslukuja.

Näiden tunnuslukujen esittelyn jälkeen syvennytään analysoimaan tarkemmin niitä piirteitä, joiden perusteella malli on luokittelun tehnyt. Näitä piirteitä tarkastelemalla teen johtopäätöksiä siitä, minkälaisia eroavaisuuksia naisista ja miehistä kertovassa uutisoinnissa on. Tekstiesimerkkejä varten kaikista alkuperäisistä virkkeistä on luotu korpukset, jotta piirteinä esiintyvistä sanoista on voitu tehdä konkordanssianalyysiä ja tuoda mukaan tekstiesimerkkejä havainnollistamaan miten kyseisiä sanoja on alkuperäisissä virkkeissä käytetty. Konkordanssianalyysiä on tehty AntConc-ohjelmalla.

Keskityn ensin Yle Uutisten ja sitten Helsingin Sanomien aineistoon ja teen sitten havaintoja aineistojen välillä olleista eroista.

6.1 Tunnusluvut, Yle Uutiset

Taulukko: Yle Uutisten aineiston tunnusluvut

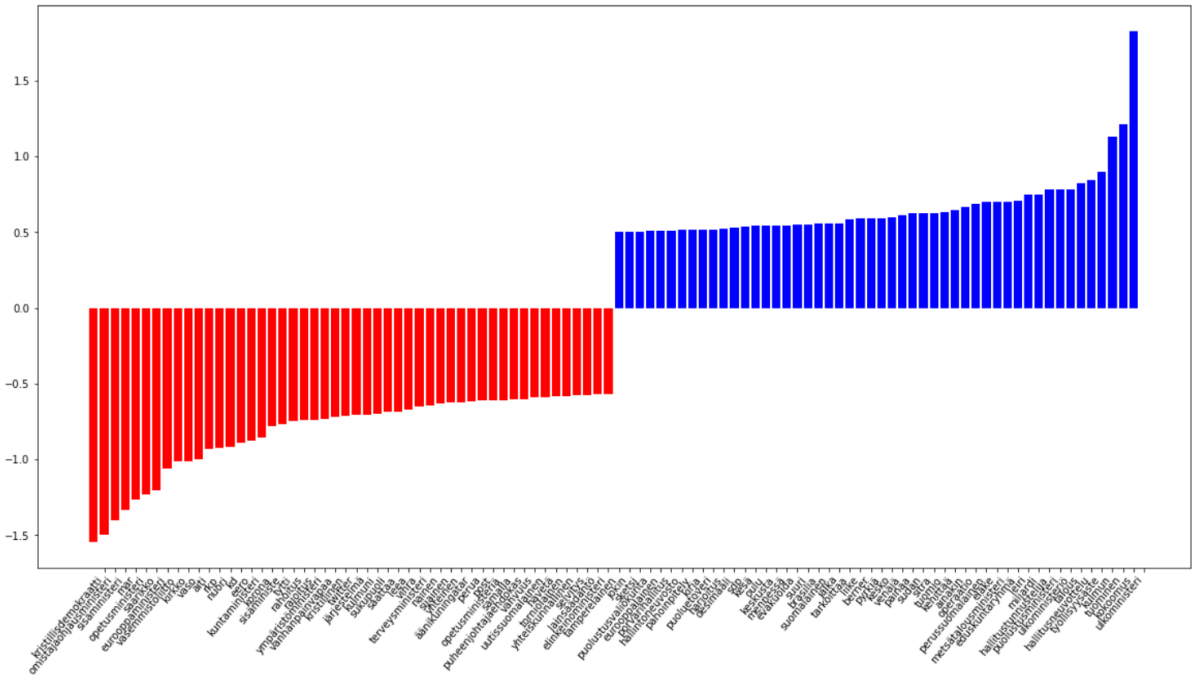
	tarkkuus	herkkyys	f1-mitta	tekstien määrä
luokka 0	0.67	0.42	0.51	1388
luokka 1	0.67	0.85	0.75	1944
		täsmällisyys:	0.67	3332

Kuten yllä olevasta kuvasta näkyy, tukivektorikone kykenee jonkin verran tunnistamaan miehistä ja naisista kertovia lauseita. Naisista kertovan tekstin se tunnistaa paremmin, f1-mitan ollessa 0.75. Kaiken kaikkiaan mallin täsmällisyys on 0.67. Testiaineistossa luokan 0 tekstejä on 1388 ja luokan 1 1944, eli jos malli ennustaisi aina luokkaa 1, jota aineistossa on eniten, täsmällisyydeksi tulisi n. 0.58, eli malli on parempi kuin tämä perustaso.

Melko alhaiseksi jäävä täsmällisyys voi viitata siihen, ettei aineistossa ole merkittävästi eroavaisuuksia miehistä ja naisista kertovien tekstien välillä. Malli ei siis kykene tekemään luokittelua hyvin, koska tekstit ovat toistensa kanssa liian samanlaisia. Tämä voisi tarkoittaa sitä, että Ylen uutisointi kansanedustajista on melko sukupuolineutraalia eikä kielessä ole

suuria eroja sukupuolten välillä. Kuitenkin, media-analyysiä tehdessä alhaisemmallakin täsmällisyydellä voidaan saada mielenkiintoisia tuloksia, kun analyysiä tehdään lisäksi laadullisesti (Heyer ym., 2006 Leavyn, 2019, mukaan).

6.2 Piirteiden tarkastelu, Yle Uutiset



Kuva 4. Piirteet kuvaajana. Isompi versio liitteissä.

Mallin jatkotarkastelua voidaan siis melko alhaisesta täsmällisyydestä huolimatta tehdä tutkimalla laadullisesti piirteitä, joiden perusteella malli on tehnyt luokittelua. Analyysin helpottamiseksi olen ryhmitellyt piirteet. Ryhmittely on yleinen tapa kvalitatiiviseen analysointiin korpusavusteisessa diskurssianalyysissä. Tässä tutkimuksessa ryhmät ovat 1. tittelit, puolueet ja tehtävät, 2. puheenaiheet ja politiikka, 3. kuvailevat sanat, 4. verbit, 5. muut ja 6. nimet. Sanoja on siis ryhmitelty sekä niiden merkityksen että sanaluokkien perusteella.

Ensimmäiseen ryhmään olen valinnut piirteet, jotka ovat selkeästi titteleitä, puolueiden nimiä tai niiden lyhenteitä tai muita tehtävänimikkeitä. Toisessa ryhmässä ovat piirteet, jotka viittaavat erilaisiin (poliittisiin) puheenaiheisiin, jotka ovat olleet mediassa esillä, sekä politiikan tekoon viittaavat piirteet, kuten esimerkiksi puoluetoveri tai vaalitulos. Kolmannessa ryhmässä ovat kuvailevat sanat, eli käytännössä adjektiivit ja adverbit, neljännessä verbit ja viidennessä sellaiset piirteet, jotka eivät ole sopineet aiempiin ryhmiin.

Kuudenteen ryhmään olen poiminut nimet, jotka ovat jääneet aineistoon esiprosessoinnista huolimatta[6].

Seuraavissa taulukoissa on näin ryhmiteltynä molemmista luokista sata suurimman korrelaatin piirrettä. Taulukoissa on myös esitetty se, montako piirrettä kussakin ryhmässä on.

Taulukko: Luokka 0 = mies, Yle Uutisten aineisto

tittelit, puolueet, tehtävät	**puheenaiheet ja politiikka**	**kuvailevat sanat**	**verbit**	**muut**	**nimet**
ulkoministeri	työllisyysaste	suomalainen	kehittää	miljardi	tuominen
kokoomus	hallitusneuvottelu	suuri	painaa	leiri	kulmun
ulkoministeriö	talous	vahvasti	pyrkiä	ainakin	berner
puolustusministeri	eduskuntaryhmä	väärä	tarkoittaa	tuomio	
hallitustunnustelija	eläke		evakuoida	jalka	
metsätalousministeri	operaatio		varoittaa	mennessä	
perussuomalainen	sitra		suorittaa	oulu	
liike	sudan		kerätä	kesä	
keskusta	venäjä		väistyä	desimaali	
sdp	kesko		osallistua	harjoitus	
hallintoneuvosto	brasilia		matkustaa	kirja	
puolustusvaliokunta	puoluetoveri			pahoinpitely	
kok	porvarihallitus			sketsi	
rikosylikomisario	eurooppalainen			jokin	
peruspalveluministeri	vaalitilaisuus			mikään	
presidentti	johnson			edellytys	
puhemiesneuvosto	virkamiesjohto			ulos	
	sotemalli			tie	
	irak			lentokenttä	
	eurooppa			facebook	
	palkkatuki			monimuotoisuus	
	hallinto			areena	
	kabuli			esimerkki	
	täysistunto			suunta	
	hallituskausi			tunnelma	
	aktiivimalli			kehitys	
	vaalit			blogi	
	nuorisosäätiö			rauma	
	nuorisojärjestö			loppu	
	äänestäjä			profiili	
	äänikuningas			ratkaisu	
				arvo	
				luonto	
				jyväskylä	
17	31	4	11	34	3

Taulukko: Luokka 1 = nainen, Yle Uutisten aineisto

tittelit, puolueet, tehtävät	**puheenaiheet ja politiikka**	**adjektiivit ja adverbit**	**verbit**	**muut**	**nimet**
kristillisdemokraatti	korona	nuori	saattaa	kirkko	mar
omistajaohjausministeri	rajoitus	yhteinen	perua	äiti	saarikko
sisäministeri	vanhanpainvapaa	torniolainen	hävetä	rahoitus	eero
opetusministeri	äänikuningatar	yhteiskunnallinen	esitellä	twitter	tytti
eurooppaministeri	puheenjohtajaehdokas	tamperelainen	sopia	järjestelmä	kulmuni

[6] Tämä voi johtua esimerkiksi siitä, että NER-vaiheessa kaikkia nimiä ei ole tunnistettu.

vasemmistoliitto vaso rkp kd kuntaministeri sisäministe ympäristöministeri kristillinen vihra terveysministeri opetusministeriä elinkeinoministeri kulttuuriministeri oikeusministeri opettaja omistajaohjaus vihreä	lainsäädäntö virkapuhelin vastaehdokas korkonatilanne poliitikko verotulo istuntokausi maahanmuuttopolitiikka	samanlainen tärkeä riittävä epätodennäköinen laaja negatiivinen hyvä kunnianhimoinen	rajata tiedottaa	sukupuoli nainen sijainen post samalla vahvuus uutissuomalainen selvitys lukio julkisuus yökerho juttu poliisi tampere suosio puhumattomuus säätiö valinta oikeusoppine jäsen puolesta päälle tviitti puuttuminen lähipäivi ohjaus testi liikunta päivämäärä kotka	leea isotalus karppinen maria mariaohisalo
22	13	13	7	35	10

6.2.1 Tittelit, puolueet, tehtävät -ryhmä

Etenkin tämän ryhmän piirteiden tarkastelu osoittaa, että käytetty luokittelija toimii niin sanotusti oikein. Mies-luokassa tästä ryhmästä löytyvät esimerkiksi ulkoministeri, puolustusministeri ja metsätalousministeri; aikana, jolta aineisto on, kaikissa näissä tehtävissä toimi mies. Ulkoministerinä toimi Pekka Haavisto, puolustusministerinä Antti Kaikkonen ja maa- ja metsätalousministerinä Jari Leppä. Nainen-luokasta löytyvät vastaavasti esimerkiksi omistajaohjausministeri ja sisäministeri; omistajaohjausministerinä toimi Tytti Tuppurainen ja sisäministereinä Maria Ohisalo ja Krista Mikkonen.

Puolueet ja niiden luokat näyttävät hyvin pitkälti seuraavaan puolueiden kansanedustajien sukupuolijakaumia. Mies-luokasta löytyvät kokoomus, keskusta, perussuomalainen, liike[7] ja sdp. Kokoomuksen kansanedustajista miehiä oli 59 prosenttia, keskustan 67 prosenttia, perussuomalaisten 71 prosenttia ja Liike Nytin 100 prosenttia. Sdp tekee poikkeuksen, sillä

[7] piirre liike viittaa suurella todennäköisyydellä Liike Nyt -puolueeseen.

puolueen kansanedustajissa miehet ovat vähemmistössä (45 prosenttia). Nainen-luokassa löytyy kristillisdemokraatteihin, vasemmistoliittoon, rkp:hen ja vihreisiin liittyviä piirteitä. Kristillisdemokraattien kansanedustajista naisia oli 60 prosenttia, vasemmistoliiton 56 prosenttia ja vihreiden 85 prosenttia. Nainen-luokassa poikkeus on rkp, jonka kansanedustajista naisia on vain 40 prosenttia.

Muut listasta löytyvät tehtävät ja sanat kuten ministeriöiden nimet osoittavat myös käytetyn metodin toimivuutta. Mielenkiintoinen huomio on se, että politiikan tekoon liittymättömät tittelit, rikosylikomisario ja opettaja, löytyvät juuri niistä luokista, joihin kyseiset tittelit stereotyyppisesti liitetään: rikosylikomisario on mies-luokassa ja opettaja luokassa nainen. Toisaalta, rikosylikomisarion päätymistä nimenomaan mies-luokkaan selittää se, että kansanedustajana toimineet Jari Kinnunen sekä Kari Tolvanen ovat ammateiltaan rikosylikomisarioita. Naiskansanedustajista ei rikosylikomisarioita löydy.

Määrällisesti tarkasteltuna naisten luokassa tähän ryhmään kuuluvia piirteitä on enemmän kuin miesten luokassa: nainen-luokassa piirteitä on 22, mies-luokassa 17.

6.2.2 Puheenaiheet ja politiikka -ryhmä

Tämän ryhmän sanoissa on nähtävissä selkeitä linkkejä edelliseen. On luonnollista, että ulkopolitiikkaan liittyvät piirteet, kuten eurooppa, eurooppalainen, sudan, venäjä ja brasilia ovat mies-luokassa, koska ulkopolitiikkaa tehtiin ulkoministeriön johdolla. Koronan myötä paljon mediassa esillä olleet rajoitukset ja korona ovat myös odotetusti nainen-luokassa, sillä Suomen koronatilanteeseen liittyvistä poliittisista päätöksistä vastasi hallitus, jonka ministereistä suuri osa oli naisia. Jotkin piirteet ovat yhdistettävissä tiettyihin henkilöihin: esimerkiksi virkapuhelin liittyy hyvin todennäköisesti Sanna Marinin virkapuhelinkohuun. Asian vahvistaa konkordanssianalyysi.

Taulukko: esimerkkejä virkkeistä, joissa sana virkapuhelin esiintyy aineistossa.

Saarikolla vain yksi puhelin käytössä Marin on kertonut, että hän ei saanut lauantai-iltana	virkapuhelimeen	lähetettyjä kehotuksia välttää kontakteja, koska hänellä oli mukana baarissa vain eduskuntapuhelin.
Marin luki valtioneuvoston	virkapuhelimeen	lähetetyt viestit vasta seuraavana aamuna eli sunnuntaina.
	Virkapuhelimeen	myöhemmin lauantai-iltana lähetetyt kehotukset välttää kontakteja jäivät Marinin mukaan saamatta, koska hänellä oli mukana baarissa vain eduskuntapuhelin.
Marinin mukaan pääministerin	virkapuhelimen	käyttö käydään tarkasti läpi valtioneuvoston kanslian ja turvallisuusyksikön kanssa.
Pääministeri Marin jätti	virkapuhelimen	tarkoituksella kotiin

Molemmissa luokissa on tässä ryhmässä nähtävissä myös yksi selkeästi sukupuolittunut piirre: mies-luokassa äänikuningas ja nainen-luokassa vastaavasti äänikuningatar. Määrällisesti ero on hyvin samanlainen kuin aiemmassa ryhmässä, mutta toisin päin: mies-luokassa piirteitä on 31, nainen-luokassa 13.

6.2.3 Adjektiivit ja adverbit -ryhmä

Mies-luokassa neljä adjektiivia tai adverbia. Aiempaan, politiikka ja puheenaiheet -ryhmään on sijoitettu tässä jaossa toinen mies-luokan piirteistä löytyvä eurooppalainen-adjektiivi sillä perusteella, että sen voi olettaa liittyvän ulkopolitiikan puheenaiheisiin. Nainen-luokassa puolestaan on 13 piirrettä tässä ryhmässä. Nuori voisi viitata siihen, että poliitikon ikä nousee uutisoinnissa esille naisten osalla miehiä todennäköisemmin. Konkordanssianalyysin perusteella näin voi olla, mutta sanaa nuori käytetään myös kirjoittaessa kansanedustajan menneisyydestä ja nuoruudesta.

Taulukko: esimerkkejä virkkeistä, joissa sana nuori esiintyy aineistossa.

Li Andersson on saavuttanut politiikassa paljon varsin	nuorella	iällä.
Kysyimme Suomelalta, Jäntiltä ja kahdelta	nuoremman	polven kaupunginvaltuutetulta näkemystä siitä, miten nuoria naispoliitikkoja kohdellaan Tampereen valtuustossa.
Ilmeisesti Kaikkonen aavisti, että hän saisi vastaehdokkaakseen itseään 14 vuotta	nuoremman	Katri Kulmunin.
Purra sanoo äänestäneensä	nuorempana	vihreitä, mutta päätyneensä lopulta perussuomalaisiin.
Puoluekokouksen aattona 32 vuotta täyttänyt Kulmuni on muutaman kuukauden	nuorempi	kuin vasemmistoliiton Li Andersson.
Hartikainen kysyi Halla-aholta, ovatko blogitekstit	nuoren	vihaisen miehen kärkkäitä kommentteja, jotka eivät edusta hänen ajatteluaan, vai seisooko hän edelleen tekstinsä takana.
Vaikka Marin on vain 3 vuotta Lindtmania nuorempi, puoluevaltuusto ilmeisesti koki nimenomaan hänet valovoimaisemmaksi	nuoren	polven edustajaksi.
Vasemmistoliiton puheenjohtaja Li Andersson esiintyi jo hyvin	nuorena	pontevasti tv:ssä isänsä kanssa.
Anna-Maja Henriksson tekee ehkä jonkinlaisen ennätyksen siinä, miten	nuorena	voi kokea jonkinlaisen poliittisen heräämisen.

6.2.4 Verbit-ryhmä

Isoimpia eroja verbeissä luokkien välillä on se, että nainen-luokan verbeistä moni kuvailee jonkin asian ilmaisua: esitellä ja tiedottaa ovat selkeitä esimerkkejä, mutta myös verbejä ajatella ja pohtia voidaan käyttää samanlaisessa yhteydessä. Mies-luokassa vastaavanlaisia verbejä on vain yksi, varoittaa.

6.2.5 Muut-ryhmä

Tästä ryhmästä löytyy toinen sukupuolittunut sana: äiti. Vastaavasti sanaa isä ei löydy mies-luokasta. Konkordanssanalyysi kuitenkin osoittaa, ettei kyseessä ole niinkään kansanedustajien äitiydestä puhuminen. Kansanedustajien äidit ovat esiintyneet naiskansanedustajista kertovissa teksteissä, ja äitiys voi liittyä myös poliittisiin kysymyksiin, kuten perhevapaisiin sekä kysymyksiin äitien ja lapsien kotiuttamiseen al-Holin leiriltä. Vaikka kyseessä ei siis aina olekaan kansanedustajan äitiys, on kansanedustajien äideistä puhuminen silti perheeseen liittyvää uutisointia.

Taulukko: esimerkkejä virkkeistä, joissa sana äiti esiintyy aineistossa.

Minja Västisen	äiti	on SDP:n kansanedustaja Piritta Rantanen.
Keskustalle on toisaalta tärkeintä, että	äitien	nykyisin käytettävissä olevat vapaat eivät vähene, perhevapaauudistusneuvotteluissakin mukana ollut Saarikko sanoo.
Eduskuntavaalien alla Saarikko muotoili Ylen vaalikoneessa kantansa (siirryt toiseen palveluun) niin, että isien osuutta perhevapaista pitää kasvattaa, mutta ei	äitien	kustannuksella.
Opetusministeri Li Andersson (vas.) toteaakin, että on	äitienkin	kotiuttaminen mahdollista – viranomaisen harkinnan mukaan.
Perhe asui noin seitsemän vuotta eri puolilla Helsinkiä, kunnes Marin muutti	äitinsä	ja tämän naisystävän kanssa 90-luvun alussa Pirkkalaan Tampereen naapuriin.
Ohisalo vietti ensimmäisen syntymäpäivänsä	äitinsä	kanssa turvakodissa.
Purra kertoo innostuneensa politiikasta ja yhteiskunnasta jo	äitinsä	kannustamana.
Ennen joulua koko oppositio teki perussuomalaisten johdolla välikysymyksen hallituksen ja erityisesti ulkoministeri Pekka Haaviston (vihr.) toimista al-Holin leirin suomalaisten lasten ja	äitien	kotiuttamisessa.

Nainen-luokasta löytyy myös piirteet sukupuoli sekä nainen, jotka voisivat viitata siihen, että sukupuolesta ylipäätään kirjoitetaan mediassa enemmän naisten kuin miesten osalta. Mies-sana ei nouse mies-luokassa sadan suurimman korrelaatin piirteen listalle.

Taulukko: esimerkkejä virkkeistä, joissa sana nainen esiintyy aineistossa.

Saarikon mukaan keskustalla ei ole naiskysymyksessä sellaista todistustaakkaa kuin esimerkiksi kokoomuksella, koska keskustalla on ollut	nainen	puheenjohtajana kahteen kertaan.
Perussuomalaisten vahvin nouseva kyky on myös	nainen,	uusmaalainen kansanedustaja Riikka Purra (siirryt toiseen palveluun).
Sanna Marin oppi nopeasti, miten voi joutua lehdistön tulitukseen, oli sitten nuori, vanha,	nainen,	mies, kokenut tai kokematon.
Viidestä hallituspuolueeesta neljän puheenjohtajana on	nainen,	ja kesäkuussa myös SDP saa naispuheenjohtajan, kun pääministeri Sanna Marin valitaan Antti Rinteen seuraajaksi.
Jos puheenjohtajasta tulee oletettu kolmiodraama, Kulmuni hyötyy sitä, että hän on asetelman ainoa	nainen.	
Tulos olisi Marinin ikäiselle	naiselle	reippaasti erinomainen.
Räsänen korosti, että avioliitto on säädetty Raamatussa	naisen	ja miehen väliseksi ja homoseksuaaliset suhteet kuvaillaan synniksi ja häpeäksi.
Marinin yli-itsevarmuus vakuuttaa ja ärsyttää SDP valitsi johtajakseen yli-itsevarman	naisen,	joka ei epäröi kehua itseään.

Kukaan meistä ei pysty valitsemaan omia vanhempiaan, mutta on ehdottomasti Suomen etu, että nämä	naiset	eivät pääse tänne takaisin, Halla-aho totesi.
– Lapsilisää voisi korottaa jo ensimmäisestä lapsesta, vähintään kahden lapsen perhe voisi saada valtion tukeman lainan ja kolme lasta synnyttäneet	naiset	voisivat vapautua vaikka kokonaan tuloverosta, Tavio sanoi puheenvuorossan.

Konkordanssianalyysi sanasta sukupuoli osoittaa, ettei äiti-sanan tapaan aina teksteissä ole kyse kansanedustajan sukupuolesta, vaan esimerkiksi translain uudistusta käsitellessä sukupuoli on ollut poliittinen puheenaihe.

Taulukko: esimerkkejä virkkeistä, joissa sana sukupuoli esiintyy aineistossa.

Kiinnostus liittyy Marinin ikään ja	sukupuoleen	sekä viiden naispuheenjohtajan valtaannousuun.
– Komissaaria, kuten muitakaan päättäjiä, ei pidä valita	sukupuolen	perusteella vaan siitä riippumatta, Halla-aho perustelee.
Marin pohdiskeli Ykkösaamussa, että	sukupuolen	juridinen vahvistaminen voisi olla mahdollista alle 18-vuotiaille, mutta sanoi kysymystä erittäin hankalaksi.
Marin hallitusohjelmaa suopeampi	sukupuolen	korjaamiselle Marin on ajanut voimakkaasti translain uudistamista, mutta ihmisoikeusjärjestö
Translakiasiassa Essayah mainitsi intersukupuolisuuden mutta ei kommentoinut sukupuolidysforiaa (ihmisen kokemusta ristiriidasta oman	sukupuolensa	ja sen kehollisen ilmaisun kanssa), vaikka tämä aihe on kysymyskokonaisuuden keskiössä.
Marin on itse kertonut, ettei että ei hänen ikänsä tai	sukupuolensa	ole määrittänyt hänen työtään poliitikkona.
Marin sanoi kuitenkin pitävänsä kiinni hallitusohjelmasta, jossa ihmisen itsemääräämisoikeus	sukupuolesta	tunnustetaan, mutta alaikäisten osalta asia ei muutu.
Kristillisdemokraattien puheenjohtaja Sari Essayah sanoi, ettei että ei	sukupuoli	voi olla ilmoitusasia.
Sanna Marin kertoo, että häneltä on on monta kertaa kysytty, miten ikä tai	sukupuoli	vaikuttavat.
Marinin	sukupuoli	näkyi myös keskusteluissa.

Miesten luokassa listalle on noussut facebook ja blogi, naisten luokassa twitter ja tviitti – tämä voi kertoa sosiaalisen median käytön eroista luokkien välillä.

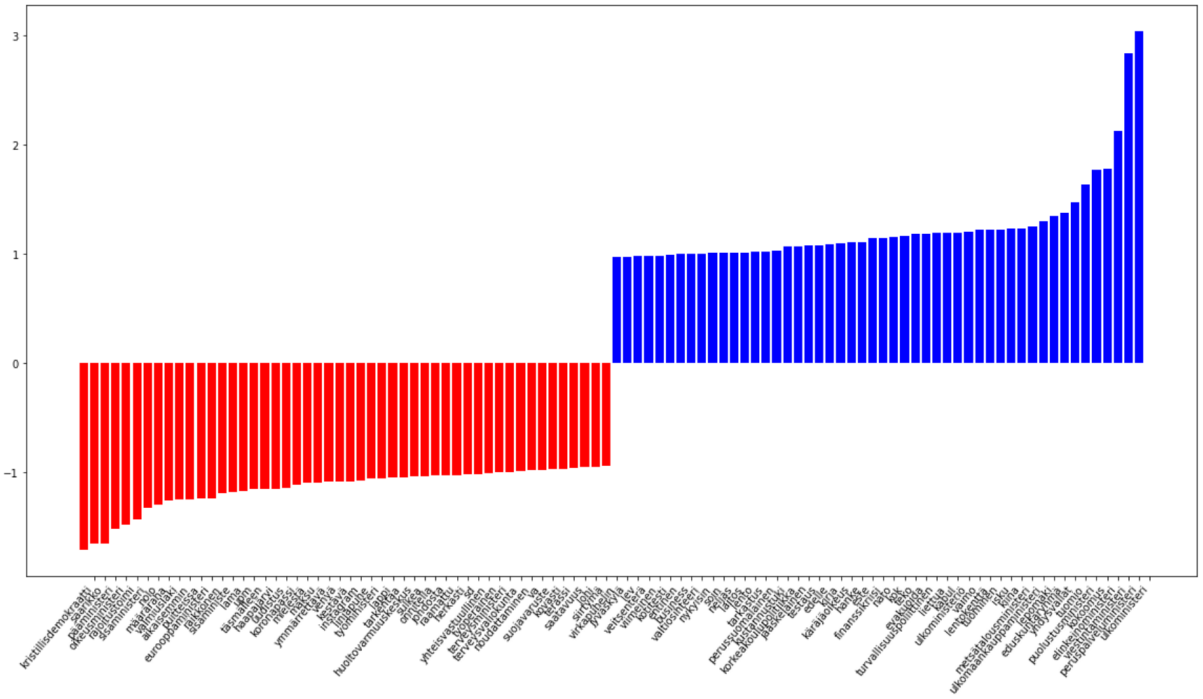
6.3 Tunnusluvut, Helsingin Sanomat

Taulukko: Helsingin Sanomien aineiston tunnusluvut

	tarkkuus	herkkyys	f1-mitta	tekstien määrä
luokka 0	0.68	0.77	0.72	1895
luokka 1	0.65	0.56	0.6	1510
		täsmällisyys:	0.67	3405

Siinä missä Yle Uutisten aineiston osalta tukivektorikone oli parempi ennustamaan luokkaa 1 = nainen, Helsingin Sanomien aineiston osalta tilanne on päinvastoin. Tämä johtunee aineistojen painotuksista: Ylen aineistossa oli enemmän luokan 1 tekstejä, Helsingin Sanomien aineistossa enemmistöluokka on puolestaan luokka 0. Kun tälle aineistolle lasketaan perustaso samalla tavalla kuin Ylen aineistolle (täsmällisyys, jos jokaisen esimerkin kohdalla ennustaan enemmistöluokkaa), perustasoksi tulee 0.56. Malli on siis tätä perustasoa parempi, kun sen täsmällisyys on 0.67. Täsmällisyys on sama kuin Yle Uutisten aineistossa.

6.4 Piirteiden tarkastelu, Helsingin Sanomien uutiset



Kuva 5. Piirteet kuvaajana. Isompi versio liitteissä.

Taulukko: Luokka 0 = mies, Helsingin Sanomien aineisto

tittelit, puolueet, tehtävät	**puheenaiheet ja politiikka**	**adjektiivit ja adverbit**	**verbit**	**muut**	**nimet**
ulkoministeri	yhdysvallat	entinen	evakuoida	tuomio	lepomäki
peruspalveluministeri	kiina	viimeinen	lopettaa	isku	tuominen
viestintäministeri	kabul	hallitsematon	lomauttaa	lentokenttä	jääskeläinen
elinkeinoministeri	liettua	nykyisin	joutua	vaimo	soin
kokoomus	turvallisuuspoliittinen	järkyttävä		teko	ojanen
puolustusministeri	nato	ääretön		este	aarnio
ulkomaankauppaministeri	finanssikriisi	valtakunnallinen		hanke	hakanen
metsätalousministeri	korkeakoulupolitiikka	kustannustehokas		käräjäoikeus	paavo
ulkoministeriö	kustannustuki			kirja	
kok	testaus			edelle	
perussuomalainen	kotitesti			tarkastus	
valtiosihteeri	afganistan			lento	
yksinyrittäjä	kannabis			laitos	
demokraatti	kotitestaaminen			neljäs	
energiaministeri	kunto			business	

suurlähetystö ulkoasianvaliokunta	hxhanke palkkatuki puolustusvoima ukraina suojamaski teollisuus traficom rokotus taksiuudistus eduskuntaryhmä presidentinvaalit eduskuntakeskustelu pääministerikausi ulkoministerikokous			veitsenterä lev jyväskylä kilpailukyky ura joku suunnitelma kysyntä yhteydenpito kymmenen todellisuus pysäyttäminen sakko korotus saavuttaminen lumi yrittäminen harja ulkopuoli	
17	29	8	4	34	8

Taulukko: Luokka 1 = nainen, Helsingin Sanomien aineisto

tittelit, puolueet, tehtävät	**puheenaiheet ja politiikka**	**adjektiivit ja adverbit**	**verbit**	**muut**	**nimet**
kristillisdemokraatti pääministeri oikeusministeri sisäministeri eurooppaministeri sisäministe työministeri sd terveysministeri terveysvaliokunta kd työnantaja juontaja eurooppaneuvosto talousvaliokunta kulttuuriministeri omistajaohjausministeri	rajoitustoimi määräraha valmiuslaki lama upm koronapassi huoltovarmuuskeskus suojavaruste terassi siirtymä virkapuhelin lisätalousarvio omistajapolitiikka maakuntavero neuvottelutulos koronakriisi altistuminen hyvinvointivaltio budjettineuvottelu ennakkovaikuttaminen elpymisväline menokehys linjapuhe kehysmenettely oikeisto haastattelutunti	nolo aikaisemmin täsmälleen ymmärrettävä kestävä herkästi yhteisvastuullinen työperäinen kovasti kohtuuton huomattavasti aidosti rento julkinen vastuullinen	venyä tarkentaa sulkea onnitella kirjata tuohtua mahtua viestittää varata hyötyä edesauttaa nauttia	puitteissa uudistus mielestä maksu instagram johdosta raamattu noudattaminen va saatavuus lohi myymälä vanhempi sunnuntaiilta juhliminen kysely kiroilu ero tavoite aikaväli hs kulku helpottaminen ohjelma vogue	saarikko räikkönen haapajärvi lappi häkämies
17	26	15	12	25	5

6.4.1 Tittelit, puolueet ja tehtävät -ryhmä

Tulokset Helsingin Sanomien aineistossa ovat hyvin samanlaiset kuin Ylen aineistossakin. Samat puolueet ja tehtävät jakautuvat samaan tapaan. Käytetty metodi toimii siis myös tässä aineistossa.

6.4.2 Puheenaiheet ja politiikka -ryhmä

Koronaan liittyvät sanat sekä myös Ylen aineistossa ollut virkapuhelin -piirre ovat esillä myös tässä aineistossa nainen-luokassa. Vastaavasti miesten luokassa tässä ryhmässä on monia ulkopolitiikkaan liittyviä piirteitä kuten maiden nimiä.

6.4.3 Adjektiivit ja adverbit -ryhmä

Tässäkin aineistossa nainen-luokassa adjektiiveja ja adverbeja on enemmän. Tämä voi kertoa siitä, että tällaisia sanoja käytetään naisista uutisissa enemmän: kieli on siis ikään kuin värikkäämpää. Naisten luokassa monet tämän ryhmän piirteet ovat sellaisia, joilla voidaan kuvata ihmistä tai hänen toimintaansa, kuten nolo, herkästi, aidosti ja rento. Konkordanssianalyysi osoittaa, ettei esimerkiksi sana ”nolo” kuitenkaan aina tarkoita juuri kansanedustajaa itseään tai hänen toimintaansa, vaikka tällaisiakin tapauksia löytyy.

Taulukko: esimerkkejä virkkeistä, joissa sana nolo esiintyy aineistossa

Kokoomuksen Ben Zyskowicz piti Saarikon syytöksiä	”noloina”.	
”Ministeri Saarikko, sanon suoraan: On	noloa,	että joudutte turvautumaan tämmöiseen tekniseen kikkailuun, kun arvostelette kokoomuksen vaihtoehtoa.
”Suomen pääministeri Sanna Marin näyttäytyy	nolona	klikkienkalastelijana.”
”Perjantai-illan kunniaksi tekivät eduskuntasalissa epäluottamuslauseen, mutta luikkivat sen tehtyään salista vähän	nolon	oloisina, eivät jääneet edes kuuntelemaan keskustelua ”, Guzenina sanoo twiitissään.
Ilta-Sanomat kirjoittaa, että koronavirukselle altistuneen pääministeri Sanna Marinin (sd) yökerhovierailu johti	noloon	tapahtumaketjuun.

Huomionarvoista on, että sana nolo esiintyy usein lainauksessa, eli sen ei voida katsoa olevan suoraan toimittajien tai median käyttämää kieltä liittyen mainittuun kansanedustajaan. Toisaalta toimituksilla on valta valita, minkälaisia sitaatteja uutisteksteihin valitaan.

Myös sanan herkästi tarkastelu näyttää, että sanaa voidaan käyttää neutraaleissa yhteyksissä, mutta myös henkilöön kohdistuvia käyttötapoja on.

Taulukko: esimerkkejä virkkeistä, joissa sana herkästi esiintyy aineistossa

Marin mukaan myös terasseilla on huolehdittava riittävistä turvaväleistä, vaikka jotkut asiantuntijat ovatkin sanoneet, että ulkona koronavirus ei leviä yhtä	herkästi.	
Tätä todistavat myös Saarikon kanssa neuvotelleet: hän puolustaa tulisieluisesti näkemystään ja saattaa tarvittaessa	herkästi	suutahtaa.
Suomela painottaa, että sote-palvelujen ulkoistaminen johtaa tutkimusten mukaan	herkästi	terveys- ja elinikäerojen kasvuun.
Tuntuu, että Kulmuni reagoi tässä tilanteessa aika	herkästi	ja erosi nopeasti.”
Marin sen sijaan hermostuu hyvinkin	herkästi.	
” Se ei siis ollut kuitti niinkään millekään ikäpolvelle, vaan ehkä ennen muuta sille, että kaikesta nykyään niin	herkästi	loukkaannutaan”, Marin sanoi.

Aidosti puolestaan vaikuttaa olevan retorinen keino, jota kenties naiskansanedustajat käyttävät miehiä useammin, mikä selittäisi sen päätymisen naisten luokan piirteisiin.

Taulukko: esimerkkejä virkkeistä, joissa sana aidosti esiintyy aineistossa

Pääministeri Sanna Marin (sd) vaikutti	aidosti	yllättyneeltä kokoomuksen puheenjohtajan Petteri Orpon kysymyksestä.
”Vain ymmärtämällä mitä saamelaiset ovat kokeneet, voimme	aidosti	löytää ratkaisuja tulevaisuuteen”, Marin sanoi ja kiitti lopuksi kolmella eri saamen kielellä.
Saarikon mukaan keskusta haluaa rakentaa	aidosti	monipaikkaisen Suomen ja esimerkiksi ajaa kaksoiskuntalaisuuden toteutumista.
Marin sanoi, että hänestä on tärkeää, että kentän toimijat ja valtionhallinnon toimijat, jotka	aidosti	myös panevat toimeen asioita, ovat mukana valmistelussa.
” Järjestelmää väärinkäytettiin, turvaa väittivät etsivänsä hekin, jotka eivät sitä	aidosti	tarvinneet”, Saarikko sanoi.
Saarikko sanoo, että hän ei	aidosti	vielä tiedä pohdinnan lopputulosta.
Purra vaikutti olevan vastauksesta	aidosti	yllättynyt.

Miesten luokan piirteet tässä ryhmässä ovat sellaisia, joilla todennäköisemmin kuvataan asioita tai ilmiöitä, kuten entinen, viimeinen, valtakunnallinen ja kustannustehokas. Konkordanssianalyysi vahvistaa tätä käsitystä.

Taulukko: esimerkkejä virkkeistä, joissa sanat entinen, viimeinen, valtakunnallinen ja kustannustehokas esiintyvät aineistossa

Näin voisi tulkita näistä äänestysnumeroista”, kommentoi	entinen	puheenjohtaja Jussi Halla-aho Tiihosen koronalinjaa.
Näin voisi tulkita näistä äänestysnumeroista”, sanoi	entinen	puheenjohtaja Jussi Halla-aho aiheesta.
Suomen	entinen	pääministeri Juha Sipilä (kesk) jättää eduskunnan tämän kauden jälkeen, kertoo Rantalakeus-lehti.
Halla-aho	viimeisessä	puheessaan puoluejohtajana: ”Nämä ovat olleet rikkaita ja palkitsevia vuosia” – kiitti nimeltä kahta ihmistä
Poistumispäätöksessä otettiin Kaikkosen mukaan huomioon se, että Yhdysvallat tarvitsee kuun	viimeisiä	päiviä omien joukkojensa poistumiseen maasta.
”Nyt on	viimeinen	hetki”, Haavisto sanoo.
Puolustusministeri Antti Kaikkonen (kesk) piti kaksi viikkoa sitten kiinnostavan puheen	valtakunnallisten	maanpuolustuskurssien avajaisissa.
Mutta	valtakunnallinen	linjaus lähtee siitä, että matalan riskistä liikunta- ja kulttuuritoimintaa voi edelleen terveysturvallisesti harjoittaa”, Kurvinen tiivisti.
	Valtakunnallinen	voitto ratkaistaan kaupungeissa, Orpo sanoi.
Tämä ei ole sellainen	kustannustehokas	tapa, johon hallitus lähtee mukaan”, Vanhanen sanoi.
Myös Tuomioja pohtii, olisiko ollut	kustannustehokkaampaa	jatkaa Hornetien käyttöikää ja katsoa, mihin suuntaan ilmapuolustusala kehittyy.
	Kustannustehokkuutta	ilmastonmuutoksen torjunnassa painotti Harjanne, mistä Purra iloitsi.

6.4.4 Verbit-ryhmä

Miesten luokassa on tässäkin ryhmässä paljon vähemmän piirteitä kuin naisten. Naisten luokassa on useita verbejä, jotka kuvaavat yksilön toimintaa, kuten tarkentaa, onnitella, tuohtua, viestittää.

Taulukko: esimerkkejä virkkeistä, joissa sanat tarkentaa, onnitella, tuohtua ja viestittää esiintyvät aineistossa

Andersson	tarkentaa	HS:lle, ettei että ei mahdollisesta koulujen avaamisesta saati sen aikataulusta ole vielä päätetty: ”Tässä vaiheessa tämä on enemmän omaa ajatteluani.
Marin	tarkensi,	ettei että ei päätös kuuluisi muutenkaan hallitukselle vaan viranomaisille.
Jutun julkaisun jälkeen Karin avustaja	tarkensi,	että kyse on 30 prosentin tavoitteesta EU-alueella.
Vasemmistoliiton puheenjohtaja Li Andersson	onnitteli	Purraa Twitterissä.
RKP:n puheenjohtaja Anna-Maja Henriksson	onnitteli	Purraa valinnasta.
Näin muiden puolueiden edustajat	onnittelevat	Riikka Purraa – ”Rikoit lasikaton”
Yleensä korostuneen rauhallinen ja hyvätuulinen Henriksson suorastaan	tuohtuu.	
Kuntakokeiluakin helpommin yleensä rauhallisen Haataisen saa hieman	tuohtumaan	muistuttamalla opposition syytöksistä etenkin valtiontaloutta kohentavien

		työllisyystoimien niukkuudesta tai niiden siirrosta.
Kokoomuksessa Saarikon kritiikistä	tuohduttiin.	
Hallitus kokoontuu maanantaina”, Marin	viestittää.	
Onko Pekonen yhä sitä mieltä, että Suomi oli varautunut koronakriisiin niin hyvin kuin julkisuuteen	viestittiin?	
” Olemme yhteistyöhakuinen EU:n jäsenmaa, mutta poikkeustilanteessakaan ei tulisi kumota maiden omaa vastuuta talouspolitiikkansa hoidosta eikä lisätä yhteisvastuullista velkaa”, Kulmuni	viestittää.	

Tämä tarkastelu jälleen osoittaa, että yksittäinenkin uutistapahtuma voi nostaa tietyn piirteen merkittäväksi: onnitella -sana liittyy hyvin vahvasti Riikka Purran valintaan Perussuomalaisten puheenjohtajaksi marraskuussa 2021 ja hänen saamistaan onnitteluista uutisoitiin ahkerasti.

6.4.5 Muut-ryhmä

Miesten luokassa on piirre vaimo, vastaavasti aviomies-sanaa ei löydy naisten luokasta. Voi kuitenkin olla, että naisten puolisoista kirjoitetaan yhtä paljon, mutta jos terminä käytetään vaihtelevasti esimerkiksi aviomiestä, puolisoa ja pelkkää mies-sanaa, niistä yksikään ei ole päätynyt sadan suurimman korrelaatin piirteen joukkoon. Kansanedustajan vaimosta puhuminen voidaan nähdä perheestä uutisoimisena samaan tapaan kuin Yle Uutisten aineiston nainen-luokan piirteissä ollut sana äiti, mutta konkordanssianalyysi osoittaa, että hyvin usein kyseessä on ollut Ano Turtiaisen liiketoiminta, jossa hänen vaimonsa on mukana. Toisaalta Antti Häkkäsen vaimosta puhuva uutisointi on melko stereotyyppistä, jossa vaimon kerrotaan tekevän enemmän kotitöitä.

Taulukko: esimerkkejä virkkeistä, joissa sana vaimo esiintyy aineistossa

Lakannut on niin ikään Ano Turtiaisen Irontime Oy, josta hän omisti rekisteritietojen mukaan 51 prosenttia ja Turtiaisen	vaimo	loput.
Kansanedustaja Turtiainen omistaa yrityksestä rekisteritietojen mukaan 49 prosenttia ja	vaimo	loput.
Häkkäsen kotona kotitöitä on pyritty jakamaan, mutta hän myöntää, että	vaimo	tekee enemmän.
Työnjako on sellainen, että Häkkänen käy kaupassa ja keittää aamu- ja iltapuurot,	vaimo	hoitaa muita kotitöitä.
Toiminnassa on sen sijaan Turtiaisen ja hänen	vaimonsa	kuntokeskusalan yritys Bobbin Oy.
Ano Turtiaisen toimivien yritysten joukkoon kuuluu myös METAL SPORT Licensing Oy, jonka hän omistaa puoliksi	vaimonsa	kanssa, rekisteritiedot vuodelta 2018 kertovat.
Turtiainen kertoo	vaimonsa	vastanneen yrityksen toiminnasta sen jälkeen, kun Turtiainen valittiin eduskuntaan.

Toiminnassa on sen sijaan Turtiaisen ja hänen	vaimonsa	kuntokeskusalan yritys Bobbin Oy.
Turtiainen kertoo	vaimonsa	vastanneen yrityksen toiminnasta sen jälkeen, kun Turtiainen valittiin eduskuntaan.
Worlds Strongest Sport Oy:n omistajia ovat viimeisimmän 2010 tiedon mukaan puoliksi Ano Turtiainen ja hänen	vaimonsa.	
Käräjäoikeuden mukaan mies lähetti Facebookin Messenger-viestipalvelun kautta Orpon puolisolle viestin, jossa hän uhkasi tappaa Orpon ja hänen	vaimonsa.	
Häkkäsen perheessä kalenterit kollataan joka viikko tarkkaan yhdessä, sillä	vaimokin	on hyvin kiireinen.
Työn ja treenien yhteydessä on solmittu ystävyyksiä ja ex-	vaimonsakin	Kosonen tapasi täällä.
Hänen seurustelusuhdettaan reviteltiin julkisuudessa sen jälkeen, kun hän oli eronnut	vaimostaan	ja kahden lapsensa äidistä Merja Vanhasesta vuonna 2006.

Tässäkin ryhmässä on tunnistettavissa tiettyihin uutistapahtumiin liittyviä piirteitä, kuten naisten luokassa oleva vogue. Tekemällä konkordanssianalyysiä Vogue-sanasta huomataan, että sana liittyy juuri Sanna Mariniin, joka esiintyi Vogue-lehden kuvissa vuonna 2020.

Taulukko: esimerkkejä virkkeistä, joissa sana Vogue esiintyy aineistossa

Ennen koronavirusepidemiaa tehdyssä britti-	Voguen	haastattelussa Marin kertaa hallitusohjelman pääkohdat: tavoitteen ilmastoneutraaliudesta, joka toteutetaan taloudellisesti kestävällä tavalla.
Kun pääministeri Sanna Marin (sd) katsoo suoraan silmiin toukokuisen	Voguen	Britannian-painoksen muotokuvasta, ei tukkaa edes huomaa.
Marin on esimerkiksi	Voguen	haastattelussa sanonut väliensä isäänsä olleen hyvin etäiset.
	Vogue	on tarvinnut Marinia paljon enemmän kuin Marin Vogueta.
Sanna Marin puhui Maria Veitolan Yökylässä-ohjelmassa siitä, miten hänen	Vogue-	lehdessä esiintymisestään ja liian paljastavasta asustaan on ryöpynnyt tyytymätöntä palautetta, kun taas aikaisempien, miespääministerien henkilökuvat on kuitattu asiallisina.
Marinin kuvan	Vogueen	on ottanut hollantilainen Anton Corbijn, eräs maailman tunnetuimmista rock-tähtien valokuvaajista.
Nyt on tarkoitus pohtia sitä, miksi Marin on	Voguen –	yli satavuotiaan vaikutusvaltaisen media-insituution – toukokuun numerossa.
Vogue on tarvinnut Marinia paljon enemmän kuin Marin	Vogueta.	

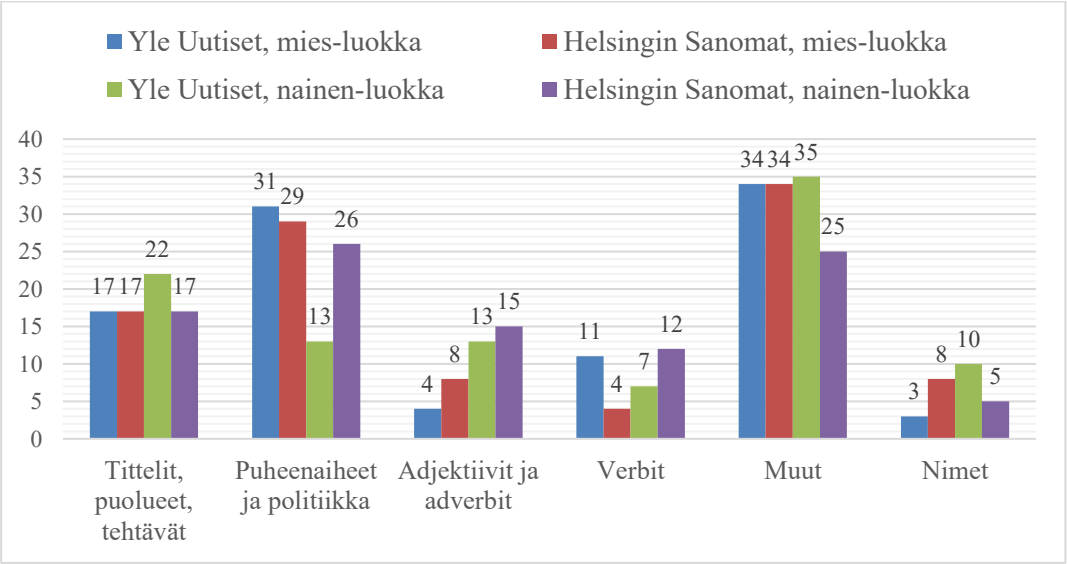
6.5 Vertailu Yle Uutisten ja Helsingin Sanomien välillä

Molempien aineistojen osalta tukivektorikoneluokittelijan täsmällisyys oli sama, 0.67. Tämä ylitti molempien perustason ja piirteiden laadullinen analyysi osoittaa luokittelijan erottaneen niin sanotusti luonnollisia eroja, kuten ministerinimikkeet, mikä vahvistaa tulkintaa siitä, että malli toimii oikein. Itse piirteissä on kuitenkin huomattavia eroja. Mies-luokan piirteitä vertailtaessa 16 piirrettä on molemmille yhteisiä, ja nainen-luokassa yhteisiä piirteitä on vain 11. Tämä voisi siis viitata siihen, että medioiden tyyleissä, aiheissa ja sananvalinnoissa on suuria eroja. Uutisten aiheiden voidaan olettaa olevan pitkälti samoja, sillä molemmat

aineistot ovat samalta aikaväliltä: Yle Uutiset ja Helsingin Sanomat ovat erittäin suurella todennäköisyydellä uutisoineet samoista politiikkaan liittyvistä aiheista. Listat yhteisistä ja eriävistä piirteistä ovat tutkimuksen liitteissä.

Listoja vertaillessa voi huomatakin sellaisia eroja, joissa on kyse sanavalinnoista. Esimerkiksi nainen-luokassa Yle Uutisten piirteistä löytyvät korkonatilanne[8] ja korona, Helsingin Sanomien piirteistä puolestaan koronakriisi sekä koronapassi. Nämä erot eivät niinkään kerro eroista sukupuolten välillä, vaan medioiden tavasta kirjoittaa samasta asiasta.

Piirteiden jakaumat eri ryhmien välillä ovat melko samanlaiset, mutta erojakin löytyy. Suurin eroavaisuus on nainen-luokassa puheenaiheet ja politiikka -ryhmän ja muut-ryhmän kohdalla.



Yle Uutisten aineiston piirteissä puheenaiheet ja politiikka -ryhmän piirteitä on nainen-luokassa selkeästi vähemmän, 13, kuin Helsingin Sanomien aineiston 26. Vastaavasti Yle Uutisten aineiston piirteissä muut-ryhmässä nainen-luokassa on 35 piirrettä ja Helsingin Sanomien aineistossa vastaavasti 25. Tämän voisi päätellä kertovan siitä, että Helsingin Sanomissa naiskansanedustajien yhteydessä kirjoitetaan enemmän politiikkaan liittyvistä aiheista ja Yle Uutisissa vähemmän. Moni muut-ryhmän piirre voi kuitenkin kertoa politiikan teosta, vaikka yksittäisen piirteen merkitys ei selvästi siihen viittaa.

[8] kyseessä mitä ilmeisimmin koronatilanne, jonka lemmatisoija on virheellisesti lemmatisoinut muotoon korkonatilanne

Molempien aineistojen osalta on nähtävissä, että mies-luokissa adjektiiveja ja adverbeja on nainen-luokkaa vähemmän. Tämä voi kertoa niin sanotusti värikkäämmästä kielenkäytöstä naiskansanedustajiin liittyvässä uutisoinnissa.

7 Johtopäätökset ja pohdinta

Käyn tässä luvussa läpi tutkimuksen keskeisimmät tulokset ja johtopäätökset ja pohdin sitten kysymyksiä, joihin ei aineiston ja sen analyysin perusteella pysty vastaamaan ja esittelen vaihtoehtoja jatkotutkimukselle.

Tämän tutkimuksen tavoitteena oli selvittää, löytyykö politiikan uutisoinnissa naisia ja miehiä koskevia eroja ja jos löytyy, millaisia erot ovat. Koostin tutkimusta varten aineiston Yle Uutisten ja Helsingin Sanomien internet-sivuilla 31.12.2019–31.12.2021 välillä julkaistuista uutisista, käsitellen uutisia niin, että lopulliseen tutkimusaineistoon päätyivät sellaiset virkkeet, joissa mainitaan kansanedustaja nimeltä. Aineisto prosessoitiin edelleen sellaiseen muotoon, että sitä voitiin analysoida käyttäen tukivektorikoneluokittelijaa.

Tukivektorikone pystyi ennustamaan, oliko virkkeessä kyse nais- vai mieskansanedustajasta laskennallista perustasoa paremmin sekä Yle Uutisten että Helsingin Sanomien aineiston osalta (Yle Uutiset perustaso 0.58, mallin täsmällisyys 0.67; Helsingin Sanomat perustaso 0.56, mallin täsmällisyys 0.67). Tämä viittaa siihen, että eroja nais- ja mieskansanedustajista kertovissa virkkeissä oli tarpeeksi siihen, että malli kykeni tekemään ennustukset perustasoa paremmin.

Jatkotarkastelussa analysoin niitä piirteitä, joiden perusteella malli teki luokittelua, keskittyen sekä nais- että mies-luokan osalta sataan suurimman korrelaatin piirteeseen, ts. sataan sellaiseen piirteeseen, jotka voimakkaimmin vaikuttivat mallin ennusteeseen. Piirteiden laadullinen analyysi yhdistettynä konkordanssitarkasteluun piirteinä olleiden sanojen esiintymisestä alkuperäisissä tutkimusaineiston virkkeissä paljasti eroavaisuuksia siinä, miten tutkittavana olleet mediat uutisoivat nais- ja mieskansanedustajista. Suuri osa piirteistä liittyi kansanedustajien tehtäviin ja sisälsi esimerkiksi ministerien titteleitä, mikä ei luonnollisesti kerro stereotypioihin nojaavista eroista vaan mies- ja naiskansanedustajien eroavista titteleistä ja tehtävistä aineiston aikavälillä. Adjektiiveja, verbejä ja muita kuin politiikkaan liittyviä sanoja tarkasteltaessa oli kuitenkin löydettävissä eroja, joita poliitikkojen tehtävät ja työ eivät selitä. Tutkimus antaa siis viitteitä siitä, että sekä Yle Uutisten että Helsingin Sanomien uutisoinnissa saattaa olla stereotypioihin nojaavia piirteitä ja että tietyt aiheet nousevat esille todennäköisemmin nais- kuin mieskansanedustajista kertovassa uutisoinnissa.

Aikaisemmissa tutkimuksissa, joissa uutisoinnin laadullisia eroja on selvitetty, on huomattu, että naispoliitikkojen ulkonäöstä, perheestä ja sukupuolesta kirjoitetaan enemmän kuin miespoliitikkojen kohdalla (van der Pas & Aaldering, 2020). Samasta ilmiöstä on viitteitä myös tämän tutkimuksen tuloksissa: piirteisiin nousivat esimerkiksi sanat äiti, sukupuoli ja nainen. Ulkonäköön viittaavia piirteitä ei löytynyt, mikä viittaisi siihen, etteivät Yle Uutiset ja Helsingin Sanomat ole uutisoineet poliitikkojen ulkonäköön liittyvistä asioista ainakaan sillä tasolla, että mahdolliset erot olisivat tulleet tässä analyysissä esiin.

Aineistona tässä tutkimuksessa oli Ylen ja Helsingin Sanomien uutisointi. Olisi mielenkiintoista tutkia, ovatko tulokset samankaltaisia muiden medioiden uutisointia analysoimalla. Samaa tutkimusmetodia voisi soveltaa myös esimerkiksi paikallisuutisiin ja siihen, miten ne käsittelevät paikallispoliitikkoja. TV-uutiset, painetut lehdet sekä radiouutiset voisivat myös antaa erilaisia tuloksia. Myös pitempi aikaväli voisi muuttaa tuloksia. Laadullinen analyysi ja konkordanssitarkastelu paljasti, että yksittäisetkin uutistapahtumat saattoivat nostaa jonkin sanan sadan suurimman korrelaatin piirteen joukkoon. Suuri osa piirteistä oli myös tehtäviä ja titteleitä. Jos aineistoa kerättäisiin pidemmältä ajalta niin, että esimerkiksi eri ministerien tehtäviä olisi hoitanut sekä naisia että miehiä, nämä ns. luonnolliset erot voisivat vaikuttaa vähemmän tuloksiin.

Aineistoa voisi tutkia myös syvemmin mitä tämän tutkimuksen puitteissa on tehty. Ovatko erot selkeämpiä ministereitä tai kansanedustajia koskevassa uutisoinnissa? Jos aineistoa ja siihen liittyviä henkilöitä tarkastellaan intersektionaalisesti, löytyisikö sukupuolen lisäksi muita erottavia tekijöitä kuten ikä, asuinpaikka tai koulutus? Onko puolueella vaikutusta siihen, miten media käsittelee poliitikkoa, entä artikkelin tai uutisen kirjoittaneen toimittajan sukupuolella?

Keskityin tässä tutkimuksessa ainoastaan niihin virkkeisiin, joissa kansanedustaja on mainittu nimeltä. Tutkimusta voisi laajentaa niin, että käsittelyssä olisivat kokonaiset artikkelit; haastavaksi tämän tekee tosin se, että politiikan uutisoinnissa samassa uutisessa on usein mainittu nimeltä useampi kuin yksi poliitikko. Feature-tyyppiset henkilöhaastattelut ovat tekstityypiltään erilaisia kuin uutisartikkelit, ja voisi siten olla mielenkiintoista selvittää, minkälaisia eroja löytyisi, jos koneoppimismallin koulutusdatana olisi henkilöhaastattelun tyyppisiä pidempiä tekstejä.

Aineiston analyysi antoi merkkejä siitä, että tarkastelussa olleiden medioiden, Yle Uutisten ja Helsingin Sanomien, tyylissä ja sanavalinnoissa voi olla suuriakin eroja. Tutkimuksen oletuksena oli, että itse uutisaiheet eivät ole medioiden välillä erilaisia, sillä aineistot on kerätty samalta ajalta ja siten voidaan olettaa, että uutisten aiheet ovat olleet samoja. Eri aiheiden painotuksissa voi kuitenkin olla eroja ja aiheiden saama palstatila voi vaihdella mediasta toiseen. Laajempien jatkotutkimusten voisi olla tarpeellista ottaa tämä tarkemmin huomioon ja pohtia, miten mahdolliset erot uutisaiheiden välillä voivat vaikuttaa tuloksiin. Uutisaiheita voisi mahdollisesti selvittää esimerkiksi topiikkimallinnuksen keinoin ja tutkia silloin myös sitä, vaikuttako uutisoitava aihe naisista ja miehistä kertovaan uutisointiin.

Tämä tutkimus keskittyi ainoastaan sanastoon: tukivektorikoneen piirteinä oli siis yksittäisiä sanoja, unigrammeja. Tutkimusta voisi laajentaa ottamalla mukaan laajempia sanojen joukkoja, kuten di- tai trigrammeja. Samaan tapaan kuin Laippalan ym. (2021) tutkimuksessa tekstilajien luokittelusta, piirteinä voisi olla myös kieliopillisia konstruktioita, jotta voitaisiin selvittää, käytetäänkö uutisteksteissä erilaisia kieliopillisia muotoja naisista ja miehistä uutisoitaessa. Tässä tutkimuksessa käytettiin myös kaikkia sanoja piirteinä: jatkotutkimuksissa voitaisiin tehdä samoin kuin Leavy (2019) ja tutkia, miten tukivektorikoneen täsmällisyys muuttuu, jos piirteiksi otetaan esimerkiksi vain tietyn sanaluokan sanat, kuten adjektiivit tai verbit.

Kuten aiemmin luvussa 2 on kerrottu, aiheen tutkimus on keskittynyt lähinnä joko kvantitatiiviseen tai kvalitatiiviseen uutisoinnin analyysiin. Tutkimus ei ole yleensä tarjonnut vastauksia siihen, mistä erot voisivat johtua. Tässäkin tutkimuksessa on keskitytty selvittämään ensin se, onko eroja, ja sitten erojen laadulliseen analyysin. Kysymys erojen syystä jää siis avoimeksi. Tämä on yleistä aiheen tutkimuksissa, kuten van der Pas ja Aaldering (2020) totesivat 90 tutkimuksen meta-analyysissään. Erojen syiden tutkiminen olisi erittäin hyödyllistä, mutta niiden selvittäminen on osoittautunut erittäin haastavaksi. Tekevätkö toimittajat ja toimitukset tietoisesti valintoja, jotka esittävät naiset ja miehet eri valossa politiikan uutisoinnissa, onko kyse alitajuntaisista stereotypioista, vai onko syy jokin aivan muu? Valitettavasti tähän ei pystytä vastaamaan tutkimalla ainoastaan sitä, onko eroja ja analysoimalla sitä, minkälaisia erot ovat.

8 Lähteet

Aaldering, L., & van der Pas, D. J. (2020). Political Leadership in the Media: Gender Bias in Leader Stereotypes during Campaign and Routine Times. *British Journal of Political Science*, *50*(3), 911–931. https://doi.org/10.1017/S0007123417000795

Aamulehden tasa-arvoiset ammattinimikkeet: tästä siinä on kyse ja tästä ei - Pääkirjoitukset - Aamulehti. (23.10.2019). Aamulehti. https://www.aamulehti.fi/paakirjoitukset/art-2000007432340.html

Barbaresi, A. (2021). Trafilatura: A Web Scraping Library and Command-Line Tool for Text Discovery and Extraction. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. https://doi.org/10.18653/v1/2021.acl-demo.15

Biecek, P., & Burzykowski, T. (2021). Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. Teoksessa *Explanatory Model Analysis*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429027192

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statist. Sci. 16(*3*)*, 199–231. https://doi.org/10.1214/SS/1009213726

Byerly, C. (2012). The geography of women and media scholarship. Teoksessa K. Ross (Toim.), *The Handbook of Gender, Sex, and Media* (ss. 3–19). Wiley-Blackwell.

Courtney, M., Breen, M., McGing, C., McMenamin, I., O'Malley, E., & Rafter, K. (2020). Underrepresenting Reality? Media Coverage of Women in Politics and Sport. *Social Science Quarterly*, *101*(4), 1282–1302. https://doi.org/10.1111/ssqu.12826

Diermeier, D., Ois Godbout, J., & Kaufmann, S. (2011). Language and Ideology in Congress. *J.Pol.S*, *42*, 31–55. https://doi.org/10.1017/S0007123411000160

Duunitorin sukupuolineutraalit ammattinimikkeet. (ei pvm.). Noudettu 15. syyskuuta 2022, osoitteesta https://duunitori.fi/sukupuolineutraalit-ammattinimikkeet

Engelberg, M. (2000). The communication of gender in Finnish. Teoksessa M. Hellinger & H. Bußmann (Toim.), *Gender across languages: The linguistic representation of women and men. Volume 2*. John Benjamins Publishing Company.

Engelberg, M. (2016). *Yleispätevä mies : Suomen kielen geneerinen, piilevä ja kieliopillistuva maskuliinisuus* [väitöskirja, Helsingin yliopisto]. Helda. https://helda.helsinki.fi/handle/10138/163013

Engelberg, M. (2018). *Miehiä ja naisihmisiä: suomen kielen seksismi ja sen purkaminen*. Tasa-arvoasiain neuvottelukunta TANE.

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, *3*, 1289–1305.

Forster, M., & Sober, E. (1994). How to Tell When Simpler, More Unified, or Less Ad Hoc Theories Will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1–35.

Guyon, I., Weston, J., & Barnhill, S. & Vapnik, V. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning 46,* 389–422. https://doi.org/10.1023/A:1012487302797

Gygax, P. M., Elmiger, D., Zufferey, S., Garnham, A., Sczesny, S., von Stockhausen, L., Braun, F., & Oakhill, J. (2019). A Language Index of Grammatical Gender Dimensions to Study the Impact of Grammatical Gender on the Way We Perceive Women and Men. *Frontiers in Psychology*, *10*, 1604. https://doi.org/10.3389/fpsyg.2019.01604

Haraldsson, A., & Wängnerud, L. (2019). The effect of media sexism on women's political ambition: evidence from a worldwide study The effect of media sexism on women's political ambition: evidence from a worldwide study. *Feminist Media Studies*, *19*(4), 525–541. https://doi.org/10.1080/14680777.2018.1468797

Hayes, D., & Lawless, J. L. (2016). *Women on the Run: Gender, Media, and Political Campaigns in a Polarized Era*. Cambridge University Press. https://doi.org/10.1017/CBO9781316336007

Heyer, G., Quasthoff, U., & Wittig, T. (2006). *Text Mining: Wissensrohstoff Text*. W3L-Verlag.

Humprecht, E., & Esser, F. (2017). A glass ceiling in the online age? Explaining the underrepresentation of women in online political news. *European Journal of Communication*, *32*(5), 439–456. https://doi.org/10.1177/0267323117720343

Joachims, T. (1998). Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Lecture Notes in Computer Science*, *1398*, 137–142. https://doi.org/10.1007/BFb0026683

Joshi, D. K., Hailu, F., & Reising, L. J. (2020). Violators, virtuous, or victims? How global newspapers represent the female member of parliament. *Feminist Media Studies*, *20*(5), 692–712. https://doi.org/10.1080/14680777.2019.1642225

JSN. (2.9.2022). *8073/A/22*. https://jsn.fi/paatos/8073-a-22/?year=2022

Julkisen sanan neuvosto. (ei pvm.). *Journalistin ohjeet*. Noudettu 1. marraskuuta 2022, osoitteesta https://jsn.fi/journalistin-ohjeet/

Kahn, K. F., & Goldenberg, E. N. (1991). Women candidates in the news: An examination of gender differences in U.S. Senate campaign coverage. *Public Opinion Quarterly*, *55*(2), 180–199. https://doi.org/10.1086/269251

Kajander, R. (29.1.2021). Näin naisten määrää nostettiin Ylen uutisten haastateltavissa – media seuraa sukupuolten tasa-arvoa yhä tarkemmin. *Yle Uutiset*. https://yle.fi/uutiset/3-11759698

Kanerva, J., Ginter, F., Miekka, N., Leino, A., & Salakoski, T. (2018). Turku Neural Parser Pipeline: An End-to-End System for the CoNLL 2018 Shared Task. *CoNLL 2018 - SIGNLL Conference on Computational Natural Language Learning, Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 133–142. https://doi.org/10.18653/V1/K18-2013

Kielitoimiston sanakirja. (ei pvm.). Noudettu 15. syyskuuta 2022, osoitteesta https://www.kielitoimistonsanakirja.fi/

Konttinen, M. (14.4.2019). Naisia nousi kansanedustajiksi historiallisen paljon – 85 prosenttia vihreiden edustajista naisia. *Yle Uutiset*. https://yle.fi/uutiset/3-10738400

Kyröläinen, A.-J., & Laippala, V. (2020). Määrällinen korpuslingvistiikka. Teoksessa M. Luodonpää-Manni, M. Hamunen, R. Konstenius, M. Miestamo, U. Nikanne, & K. Sinnemäki (Toim.), *Kielentukimuksen menetelmiä I-IV* (ss. 487–524). SKS Finnish Literature Society. https://doi.org/10.21435/SKST.1457

Laippala, V., Egbert, J., Biber, D., & Kyröläinen, A. J. (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation*, *55*(3), 757–788. https://doi.org/10.1007/S10579-020-09519-Z

Leavy, S. (2019). Uncovering Gender Bias in Media Coverage of Politicians with Machine Learning. *Digital Scholarship in the Humanities 34*(1), 48–63. https://doi.org/10.1093/llc/fqy005

Lehtonen, J. (2018). *Uutisia miehistä ja rahasta. Politiikan sukupuolittunut työnjako Yleisradion televisiouutisissa 1988–2018*.

Luhtakallio, E. (2016). Visuaalinen julkisuus ja sukupuolten representaatio. Teoksessa M. Husso & R. Heiskala (Toim.), *Sukupuolikysymys*. Gaudeamus.

Lundelin, K. (21.5.2021). Sanna Marin kaipaa vapaampaa elämää ja pois valokeilasta. *Seura*. https://seura.fi/viihde/julkkikset/paaministeri-sanna-marin-kaipaa-pois-valokeilasta/

Luoma, J., Oinonen, M., Pyykönen, M., Laippala, V., & Pyysalo, S. (2020). A Broad-coverage Corpus for Finnish Named Entity Recognition. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4615–4624.

Mills, S. (2008). *Language and Sexism*. Cambridge University Press. https://doi.org/10.1017/CBO9780511755033

Partington, A., Duguid, A., & Taylor, C. (2013). *Patterns and meanings in discourse: Theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.55

Retriever. (2021). *Sukupuolten tasa-arvo urheilun ja politiikan uutisoinnissa vuonna 2020*. Noudettu 20. marraskuuta 2021, osoitteesta https://www.sttinfo.fi/data/attachments/00063/3a3b1111-4572-41f3-8b50-4635ed816a38.pdf

Reunanen, E., Alanne, N., Räty, R., Nousuniemi, N., Harakka, T., Nuorgam, E., Toivanen, J., & Luoma-Aho, V. (2021). *Uutismedia verkossa 2021. Reuters-instituutin Digital News Report - Suomen maaraportti*. Tampereen yliopisto. https://urn.fi/URN:ISBN:978-952-03-2023-2

Rohrbach, T., Fiechtner, S., Schönhagen, P., & Puppis, M. (2020). More Than Just Gender: Exploring Contextual Influences on Media Bias of Political Candidates. *The International Journal of Press/Politics*, *2020*(4), 692–711. https://doi.org/10.1177/1940161220912694

Ruokangas, P., Juntti, M.-L., Kajander, R., & Konttinen, M. (2020, tammikuuta 29). Suomen ensimmäinen koronavirustartunta varmistui, THL: Tapaus valitettava, mutta ei odottamaton – Yle seurasi hetki hetkeltä. *Yle Uutiset*. https://yle.fi/uutiset/3-11181717

Salminen, H. (2018). *Erityistä naiseutta, normatiivista mieheyttä? : politiikan journalismin sukupuolittuminen Ylen Vaaligallerian vaalitenttivideoissa eduskuntavaaleissa 2015* [pro gradu -työ, Tampereen yliopisto]. Trepo. https://trepo.tuni.fi/handle/10024/102708

Argamon, S., Goulain J., Horton, R. & Olsen, M. (2009). Vive la Différence! Text Mining Gender Difference in French Literature. *Digital Humanities Quarterly*, *3*(2).

Shmueli, G. (2010). To Explain or to Predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Siukola, R., Kuusipalo, J., & Haapea, K. (2020). *Sukupuolella väliä eduskunnassa? - Sukupuolten tasa-arvo eduskuntaryhmien ja valiokuntien toiminnassa*. THL. http://urn.fi/URN:ISBN:978-952-343-457-8

Siukola, R., & Teräsaho, M. (2021, syyskuuta 8). Sukupuolella väliä kielessä? – THL luopuu sukupuolitetuista ammattinimikkeistä [blogikirjoitus]. *THL-blogi*.

https://blogi.thl.fi/sukupuolella-valia-kielessa-thl-luopuu-sukupuolitetuista-ammattinimikkeista/

Suero Montero, C., Munezero, M., & Kakkonen, T. (2014). Investigating the role of emotion-based features in author gender classification of text. *International Conference on Intelligent Text Processing and Computational Linguistics*, *8404*, 98–114. https://doi.org/10.1007/978-3-642-54903-8_9

Talvitie, E. (2014). *Keitäs tyttö kahvia : naisia politiikan portailla*. WSOY.

Thomas, M., Harell, A., Rijkhoff, S. A. M., & Gosselin, T. (2021). Gendered News Coverage and Women as Heads of Government. *Political Communication*, *38*(4), 388–406. https://doi.org/10.1080/10584609.2020.1784326

Trimble, L., Curtin, J., Wagner, A., Auer, M., Woodman, V. K. G., & Owens, B. (2021). Gender novelty and personalized news coverage in Australia and Canada. *International Political Science Review*, *42*(2), 164–178. https://doi.org/10.1177/0192512119876083

Valtioneuvosto. (ei pvm.). *Marinin hallitus*. Noudettu 18. lokakuuta 2022, osoitteesta https://valtioneuvosto.fi/marinin-hallitus

Valtioneuvosto. (25.3.2020). *Uudellemaalle liikkumisrajoituksia - Hallitus päätti uusista lisätoimista koronaepidemian leviämisen estämiseksi*. https://valtioneuvosto.fi/-//10616/uudellemaalle-liikkumisrajoituksia-hallitus-paatti-uusista-lisatoimista-koronaepidemian-leviamisen-estamiseksi

van der Pas, D. J., & Aaldering, L. (2020). Gender Differences in Political Media Coverage: A Meta-Analysis. *Journal of Communication*, *70*(1), 114–143. https://doi.org/10.1093/joc/jqz046

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer New York. https://doi.org/10.1007/978-1-4757-2440-0

Vasantola, S. (7.3.2021). Ikuinen kolmannes. *Helsingin Sanomat*. https://www.hs.fi/sunnuntai/art-2000007842678.html

Vidal-Correa, F. (2020). Media coverage of women in politics: Mexican local politicians on campaign. *The Journal of International Communication*, *26*(1), 1–19. https://doi.org/10.1080/13216597.2020.1736599

Yle. (ei pvm.). *Eduskuntavaalit 2019*. Noudettu 18. lokakuuta 2022, osoitteesta https://vaalit.yle.fi/ev2019/fi

9 Liitteet

Liite 1: Luokittelussa käytetty lista kansanedustajista

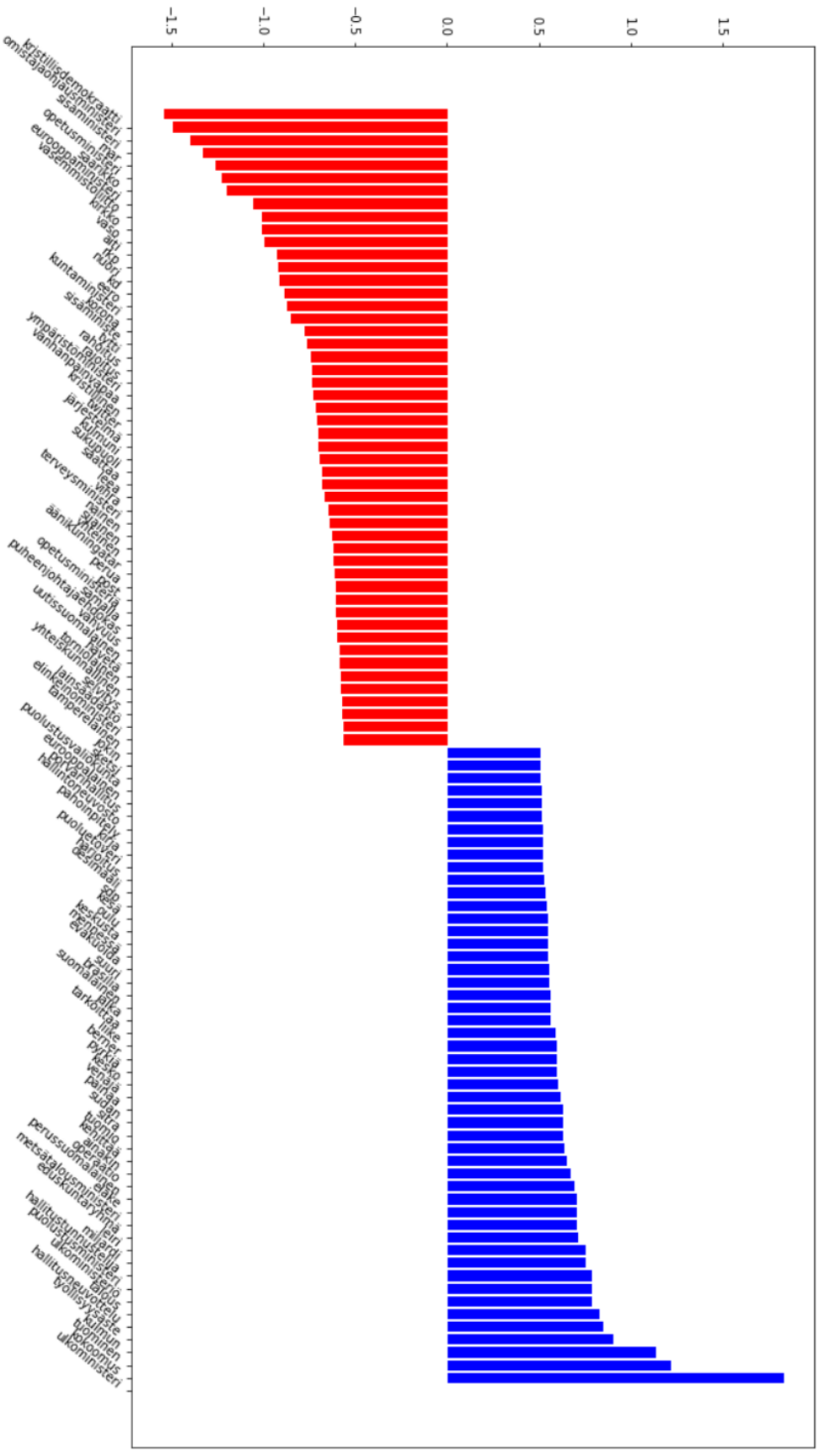
Nimi	Leima
Anders Adlercreutz	M
Pekka Aittakumpu	M
Outi Alanko-Kahiluoto	F
Li Andersson	F
Sanna Antikainen	F
Marko Asell	M
Heikki Autto	M
Kim Berg	M
Sandra Bergqvist	F
Eva Biaudet	F
Thomas Blomqvist	M
Juho Eerola	M
Markku Eestilä	M
Tiina Elo	F
Ritva Elomaa	F
Eeva-Johanna Eloranta	F
Seppo Eskelinen	M
Sari Essayah	F
Tarja Filatov	F
Bella Forsgrén	F
Sanni Grahn-Laasonen	F
Jukka Gustafsson	M
Maria Guzenina	F
Tuula Haatainen	F
Pekka Haavisto	M
Jussi Halla-aho	M
Timo Harakka	M
Atte Harjanne	M
Harry Harkimo	M
Satu Hassi	F
Hannakaisa Heikkinen	F
Janne Heikkinen	M
Timo Heinonen	M
Eveliina Heinäluoma	F
Anna-Maja Henriksson	F
Hanna Holopainen	F
Mari Holopainen	F
Veronika Honkasalo	F
Petri Honkonen	M
Inka Hopsu	F
Hannu Hoskonen	M
Petri Huru	M
Hanna Huttunen	F
Saara Hyrkkö	F
Antti Häkkänen	M
Katja Hänninen	F
Olli Immonen	M
Kalle Jokinen	M
Vilhelm Junnila	M
Kaisa Juuso	F
Arja Juvonen	F
Heli Järvinen	F
Antti Kaikkonen	M
Atte Kaleva	M
Eeva Kalli	F

Anne Kalmari	F
Ilkka Kanerva	M
Toimi Kankaanniemi	M
Emma Kari	F
Mika Kari	M
Pia Kauma	F
Ville Kaunisto	M
Juho Kautto	M
Hilkka Kemppi	F
Pihla Keto-Huovinen	F
Tuomas Kettunen	M
Anneli Kiljunen	F
Kimmo Kiljunen	M
Marko Kilpi	M
Jari Kinnunen	M
Mikko Kinnunen	M
Krista Kiuru	F
Pauli Kiuru	M
Mai Kivelä	F
Esko Kiviranta	M
Pasi Kivisaari	M
Anna Kontula	F
Ari Koponen	M
Noora Koponen	F
Jukka Kopra	M
Jari Koskela	M
Johannes Koskinen	M
Hanna Kosonen	F
Jouni Kotiaho	M
Terhi Koulumies	F
Katri Kulmuni	F
Antti Kurvinen	M
Johan Kvarnström	M
Merja Kyllönen	F
Suna Kymäläinen	F
Mikko Kärnä	M
Joonas Könttä	M
Sheikki Laakso	M
Mia Laiho	F
Antero Laukkanen	M
Rami Lehto	M
Jari Leppä	M
Aki Lindén	M
Antti Lindtman	M
Mika Lintilä	M
Markus Lohi	M
Pia Lohikoski	F
Mikko Lundén	M
Mats Löfström	M
Niina Malm	F
Sanna Marin	F
Matias Marttinen	M
Hanna-Leena Mattila	F
Leena Meri	F
Krista Mikkonen	F
Sari Multala	F
Markus Mustajärvi	M
Kai Mykkänen	M
Jari Myllykoski	M
Juha Mäenpää	M
Jani Mäkelä	M
Riitta Mäkinen	F
Merja Mäkisalo-Ropponen	F

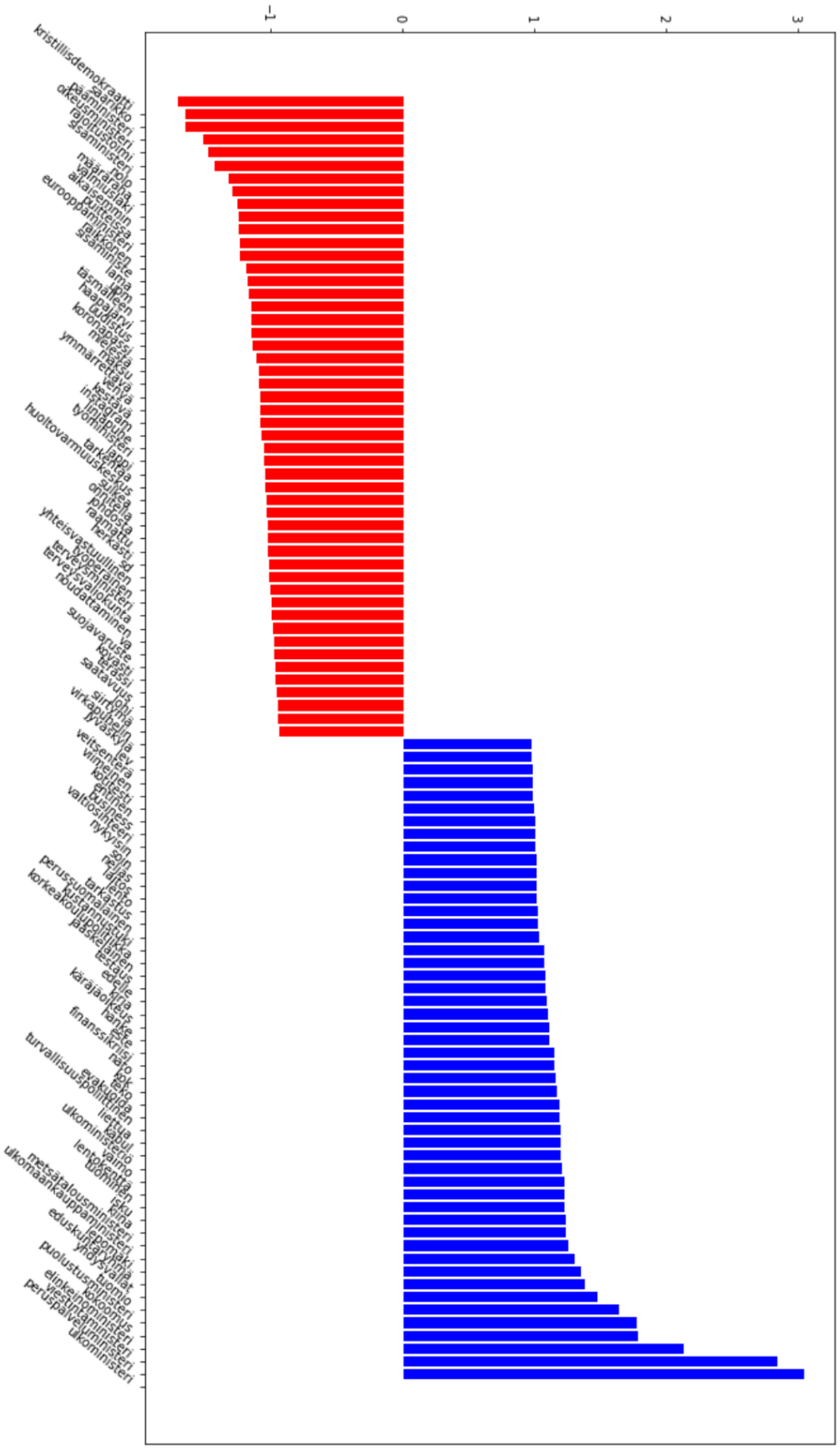
Jukka Mäkynen	M
Matias Mäkynen	M
Veijo Niemi	M
Mika Niikko	M
Anders Norrback	M
Ilmari Nurminen	M
Maria Ohisalo	F
Johanna Ojala-Niemelä	F
Mikko Ollikainen	M
Petteri Orpo	M
Jouni Ovaska	M
Sirpa Paatero	F
Tom Packalén	M
Aino-Kaisa Pekonen	F
Jaana Pelkonen	F
Mauri Peltokangas	M
Pirkka-Pekka Petelius	M
Raimo Piirainen	M
Arto Pirttilahti	M
Jenni Pitko	F
Sakari Puisto	M
Riikka Purra	F
Juha Pylväs	M
Lulu Ranne	F
Mari Rantanen	F
Piritta Rantanen	F
Veronica Rehn-Kivi	F
Minna Reijonen	F
Antti Rinne	M
Paula Risikko	F
Jari Ronkainen	M
Wille Rydman	M
Päivi Räsänen	F
Annika Saarikko	F
Suldaan Said Ahmed	M
Kristiina Salonen	F
Janne Sankelo	M
Jussi Saramo	M
Hanna Sarkkinen	F
Sari Sarkomaa	F
Arto Satonen	M
Sami Savio	F
Mikko Savola	M
Matti Semi	M
Jenna Simula	F
Juha Sipilä	M
Saara-Sofia Sirén	F
Ruut Sjöblom	F
Ville Skinnari	M
Riikka Slunga-Poutsalo	F
Mirka Soinikoski	F
Joakim Strand	M
Iiris Suomela	F
Hussein al-Taee	M
Katja Taimela	F
Mari-Leena Talvitie	F
Sari Tanus	F
Ville Tavio	M
Kari Tolvanen	M
Ari Torniainen	M
Erkki Tuomioja	M
Tytti Tuppurainen	F
Ano Turtiainen	M

Sebastian Tynkkynen	M
Veikko Vallin	M
Sinuhe Wallinheimo	M
Elina Valtonen	F
Matti Vanhanen	M
Anu Vehviläinen	F
Paula Werning	F
Heikki Vestman	M
Jussi Wihonen	M
Pia Viitanen	F
Sofia Vikman	F
Heidi Viljanen	F
Anne-Mari Virolainen	F
Sofia Virta	F
Ville Vähämäki	M
Tuula Väätäinen	F
Johannes Yrttiaho	M
Ben Zyskowicz	M
Peter Östman	M

Liite 2: Piirteet kuvaajana, Yle Uutiset



Liite 3: Piirteet kuvaajana, Helsingin Sanomat



Liite 4: Yhteiset ja eriävät piirteet aineistoissa

Mies-luokka

Yhteiset piirteet	Vain Yle Uutisten aineistossa	Vain Helsingin Sanomien aineistossa
eduskuntaryhmä	ainakin	aarnio
evakuoida	aktiivimalli	afganistan
jyväskylä	areena	business
kirja	arvo	demokraatti
kok	berner	edelle
kokoomus	blogi	eduskuntakeskustelu
lentokenttä	brasilia	elinkeinoministeri
metsätalousministeri	desimaali	energiaministeri
palkkatuki	edellytys	entinen
peruspalveluministeri	eläke	este
perussuomalainen	esimerkki	finanssikriisi
puolustusministeri	eurooppa	hakanen
tuominen	eurooppalainen	hallitsematon
tuomio	facebook	hanke
ulkoministeri	hallinto	harja
ulkoministeriö	hallintoneuvosto	hxhanke
	hallituskausi	isku
	hallitusneuvottelu	joku
	hallitustunnustelija	joutua
	harjoitus	järkyttävä
	irak	jääskeläinen
	jalka	kabul
	johnson	kannabis
	jokin	kiina
	kabuli	kilpailukyky
	kehittää	korkeakoulupolitiikka
	kehitys	korotus
	kerätä	kotitestaaminen
	kesko	kotitesti
	keskusta	kunto
	kesä	kustannustehokas
	kulmun	kustannustuki
	leiri	kymmenen
	liike	kysyntä
	loppu	käräjäoikeus
	luonto	laitos
	matkustaa	lento
	mennessä	lepomäki
	mikään	lev
	miljardi	liettua
	monimuotoisuus	lomauttaa
	nuorisojärjestö	lopettaa
	nuorisosäätiö	lumi
	operaatio	nato
	osallistua	neljäs
	oulu	nykyisin
	pahoinpitely	ojanen
	painaa	paavo
	porvarihallitus	presidentinvaalit
	presidentti	puolustusvoima
	profiili	pysäyttäminen
	puhemiesneuvosto	pääministerikausi
	puoluetoveri	rokotus
	puolustusvaliokunta	saavuttaminen
	pyrkiä	sakko
	ratkaisu	soin
	rauma	suojamaski
	rikosylikomisario	suunnitelma

	sdp sitra sketsi sotemalli sudan suomalainen suorittaa suunta suuri talous tarkoittaa tie tunnelma työllisyysaste täysistunto ulos vaalit vaalitilaisuus vahvasti varoittaa venäjä virkamiesjohto väistyä väärä äänestäjä äänikuningas	suurlähetystö taksiuudistus tarkastus teko teollisuus testaus todellisuus traficom turvallisuuspoliittinen ukraina ulkoasianvaliokunta ulkomaankauppaministeri ulkoministerikokous ulkopuoli ura vaimo valtakunnallinen valtiosihteeri veitsenterä viestintäministeri viimeinen yhdysvallat yhteydenpito yksinyrittäjä yrittäminen ääretön

Nainen-luokka

Yhteiset piirteet	Vain Yle Uutisten aineistossa	Vain Helsingin Sanomien aineistossa
eurooppaministeri kd kristillisdemokraatti kulttuuriministeri oikeusministeri omistajaohjausministeri saarikko sisäministe sisäministeri terveysministeri virkapuhelin	eero elinkeinoministeri epätodennäköinen esitellä hyvä hävetä isotalus istuntokausi julkisuus juttu järjestelmä jäsen karppinen kirkko korkonatilanne korona kotka kristillinen kulmuni kunnianhimoinen kuntaministeri laaja lainsäädäntö leea liikunta lukio lähipäivi maahanmuuttopolitiikka mar maria mariaohisalo nainen	aidosti aikaisemmin aikaväli altistuminen budjettineuvottelu edesauttaa elpymisväline ennakkovaikuttaminen ero eurooppaneuvosto haapajärvi haastattelutunti helpottaminen herkästi hs huoltovarmuuskeskus huomattavasti hyvinvointivaltio hyötyä häkämies instagram johdosta juhliminen julkinen juontaja kehysmenettely kestävä kirjata kiroilu kohtuuton koronakriisi koronapassi

	negatiivinen	kovasti
	nuori	kulku
	ohjaus	kysely
	oikeusoppine	lama
	omistajaohjaus	lappi
	opettaja	linjapuhe
	opetusministeri	lisätalousarvio
	opetusministeriä	lohi
	perua	maakuntavero
	poliisi	mahtua
	poliitikko	maksu
	post	menokehys
	puheenjohtajaehdokas	mielestä
	puhumattomuus	myymälä
	puolesta	määräraha
	puuttuminen	nauttia
	päivämäärä	neuvottelutulos
	päälle	nolo
	rahoitus	noudattaminen
	rajata	ohjelma
	rajoitus	oikeisto
	riittävä	omistajapolitiikka
	rkp	onnitella
	saattaa	puitteissa
	samalla	pääministeri
	samanlainen	raamattu
	selvitys	rajoitustoimi
	sijainen	rento
	sopia	räikkönen
	sukupuoli	saatavuus
	suosio	sd
	säätiö	siirtymä
	tampere	sulkea
	tamperelainen	sunnuntaiilta
	testi	suojavaruste
	tiedottaa	talousvaliokunta
	torniolainen	tarkentaa
	tviitti	tavoite
	twitter	terassi
	tytti	terveysvaliokunta
	tärkeä	tuohtua
	uutissuomalainen	työministeri
	vahvuus	työnantaja
	valinta	työperäinen
	vanhanpainvapaa	täsmälleen
	vasemmistoliitto	upm
	vaso	uudistus
	vastaehdokas	va
	verotulo	valmiuslaki
	vihra	vanhempi
	vihreä	varata
	yhteinen	vastuullinen
	yhteiskunnallinen	venyä
	ympäristöministeri	viestittää
	yökerho	vogue
	äiti	yhteisvastuullinen
	äänikuningatar	ymmärrettävä