KATSAUS

Kimmo Elo

Koeporauksia laskennallisin menetelmin kansanedustajien eduskuntapuheeseen vuosina 1991–1999 FinParl-aineiston avulla

ABSTRAKTI/ABSTRACT

Tässä katsausartikkelissa esitellään pienten analyyttisten koeporausten avulla eduskunnan täysistuntopuheet vuodesta 1907 lähtien koneluettavassa muodossa tarjoavan FinParl-aineiston mahdollisuuksia digitaaliselle parlamenttitutkimukselle. Analyyseissä havainnollistetaan laskennallisten menetelmien käyttöä isojen aineistojen eksploratiivisen eli uutta kartoittavan analyysin välineinä ja tarkastellaan saatuja tuloksia suhteessa parlamenttitutkimuksen kenttään. Analyyseissä fokusoidutaan erityisesti kansanedustajien täysistuntopuheiden sanastoon, jonka avulla analysoidaan hallitus-oppositio-asetelman sekä edustajien puoluetaustan vaikutusta täysistuntopuhuntaan. Lisäksi havainnollistetaan sanapohjaisen analyysin mahdollisuuksia tutkittaessa puhujien semanttista läheisyyttä sekä täysistuntokeskusteluissa käsiteltyjä teemoja. Kokonaisuutena esitetyt analyysit osoittavat laskennallisten menetelmien avaavan kiinnostavia mahdollisuuksia juuri laajojen tekstikokonaisuuksien tutkimukselle ja siten mahdollistavat esimerkiksi pitkän aikavälin muutosten analysoinnin parlamenttipuheelle tyypillisten konventioiden, piilevien argumentaatiorakenteiden tai temaattisen syklisyyden tutkimiseksi. Laskennallisia menetelmiä ei tulisikaan nähdä kilpailija muille menetelmille, vaan tutkimusvälineistöä rikastuttavana tutkijan työkaluna.

Eduskuntapuhe, laskennalliset menetelmät, digitaaliset ihmistieteet

VTT, dosentti Kimmo Elo, Turun yliopiston eduskuntatutkimuksen keskus, Eurooppa-tutkimuksen erikoistutkija, kimmo.elo@utu.fi

Johdanto

Teknisen kehityksen edistämä digitalisaatio on 2000-luvulla näkynyt entistä voimakkaammin myös yhteiskuntien perustoiminnoissa, kun sekä yksityistä että julkista toimintaa on enenevässä ja kiihtyvässä määrin muutettu tietoverkkopohjaisiksi, sähköisiksi palveluiksi. Siinä missä tämä, digitalisaatioksi nimetty ilmiö jo itsessään muodostaa ihmistieteille tutkimuskohteen yhteiskunnallisen muutoksen näkökulmasta, vaikuttaa digitalisaatio voimakkaasti myös käytännön tutkimustyöhön. Tästä kertoo osaltaan digitaalisten ihmistieteiden nimellä kulkevan tutkimusalueen vakiintuminen, minkä tutkimusalueen piirissä sovelletaan alun perin usein tietojenkäsittelytieteiden ja datatieteiden piirissä kehitettyjä menetelmiä ja työkaluja ihmistieteille ominaisten ilmiöiden ja kysymysten tutkimuksessa. Näiden työkalujen käytön laajeneminen ihmistieteisiin on osaltaan mahdollistanut digitalisaation mukanaan tuoma sähköisten aineistojen kasvu. Kyse ei ole pelkästään digitoituvan yhteiskunnan tuottamista sähkösyntyisistä aineistoista, joiden tarjonta on viimeisen vuosikymmenen aikana kasvanut lähes eksponentiaalisesti. Merkittävä rooli on myös digitoiduilla aineistoilla eli sellaisilla sähköiseen muotoon siirretyillä aineistoilla, jotka alun perin on tallennettu ei-sähköisesti esimerkiksi paperilla, valokuvina tai maalauksina. Kun tähän kehitykseen liittyy myös sähköisten aineistopalvelujen voimakas kehitys ja erilaisten aineistojen avaaminen vapaaseen käyttöön, pystyy tutkija nykyisin käyttämään omalta tietokoneeltaan laajasti sellaisia aineistoja, joiden hyödyntäminen vaati aiemmin runsaasti matkustelua tai joiden käyttö ei ollut lainkaan mahdollista. Digitointi on kuitenkin hidasta ja resurssi-intensiivistä, mistä syystä aineistojen määrän lisääntyminen ei tarkoita aineiston tarjontaan aiemmin liittyneiden vinoumien poistumista. Digitointihankkeet kohdistuvat lähes poikkeuksetta valikoituihin aineistokokonaisuuksiin, aineistojen löytäminen edellyttää nykyisin vähintäänkin hyviä tiedonhankintataitoja ja sähköisten aineistojen tutkiminen digitaalisin menetelmin ja välinein vaatii tutkijalta laskennallisten menetelmien hallintaa.

Digitaaliset ihmistieteet näyttävätkin muuttavan ihmistieteellistä tutkimustoimintaa suuntaan, jota tässä kutsun luonnontieteellistymiseksi, ja jolla tarkoitan muutosta pois tutkijakeskeisestä tieteenteosta kohti tutkimusryhmävetoista tutkimustoimintaa, jossa tutkimustyötä tehdään entistä useammin ryhmissä, jonka jäsenistä kukin tuo ryhmään oman, usein kapean osaamisensa. Erityisen selkeästi tämä näyttäisi koskevan juuri menetelmäosaamista digitaalisten tutkimusmenetelmien osalta, koska näiden opettaminen ihmistieteiden piirissä on edelleen varsin vakiintumatonta ja satunnaista. Kun digitaalisten menetelmien oppiminen vaatii myös erittäin hyviä tietoteknisiä valmiuksia, näitä valmiuksia omaamaton ihmistieteilijä, joka haluaisi soveltaa digitaalisia menetelmiä, on käytännössä pakotettu yhteistyöhön näitä menetelmiä hallitsevien tutkijoiden kanssa. Monissa tapauksissa tämä tarkoittaa myös erilaisten tutkimuskonseptien ja ontologisten näkemysten yhteensovittamista, jossa suurin haaste näyttäisi liittyvän siihen, miten pystytään pitämään kiinni ihmistieteelliselle tutkimukselle ominaisista tutkimuskysymyksistä ja näkökulmista ja välttämään tutkimuskysymysten uudelleenmuotoilu teknis-menetelmällisten välineiden ohjaamana.

Tässä katsausartikkelissa esitellään lyhyen esimerkkianalyysin tukemana Turun yliopiston eduskuntatutkimuksen keskuksen, Aalto-yliopiston semanttisen laskennan tutkimusryhmän ja Helsingin yliopiston HELDIG-tutkimuskeskuksen yhteisessä konsortiohankkeessa muodostetun FinParl-aineistokorpuksen mahdollisuuksia ja rajoja erityisesti laskennallisten ihmistieteiden näkökulmasta. Seuraavassa luvussa tarkemmin esiteltävän aineiston ytimen muodostavat eduskunnan täysistuntopöytäkirjat vuodesta 1907 alkaen, joiden pöytäkirjojen kokotekstit ovat aineiston kautta tutkijoiden käytössä. FinParl-aineiston tekee ainutlaatuiseksi juuri sen laajuus ja kumuloituvuus eli aineistoa päivitetään säännöllisesti lisäämällä siihen uusimmat eduskunnan täysistuntopuheenvuorot. Juuri jälkimmäinen erottaa FinParl-aineiston esimerkiksi Kielipankin (www.kielipankki.fi) aineistoihin

sisältyvästä korpuksesta, joka kattaa vain aikavälin 10.09.2008–01.07.2016. FinParl-aineiston ansiosta tutkijoilla tulee siis olemaan käytössään aineistokorpus, joka ei ulotu ainoastaan aivan suomalaisen parlamentarismin alkuhetkiin saakka, vaan myös sisältää jatkuvasti eteenpäin siirtyvään nykyhetkeen liittyvät keskustelut.

Katsausartikkelillani on kaksi keskeistä tavoitetta. Ensimmäiseksi, artikkeli esittelee vuoden 2023 avoimena datana tarjolle tulevan FinParl-aineiston kuvaamalla aineiston muodostamisprosessin sekä aineiston tekniset ominaisuudet. Tämä, ensimmäisessä pääluvussa suoritettava aineiston esittely käsittelee myös laajempaa kysymystä digitoidun ja koneluettavan aineiston eroista sekä tämän erottelun käytännön implikaatioista. Toiseksi, artikkeli havainnollistaa aineiston avaamia mahdollisuuksia yhteiskuntatieteelliselle tutkimukselle esittelemällä etäluentaan perustuvia tuloksia kansanedustajien eduskuntapuheesta vuosina 1991–1999. Tämä, sinänsä varsin suppea analyysi pyrkii havainnollistamaan laskennallisiin menetelmiin perustuvan tutkimusprosessin käytännön toteuttamista ja tuomaan esille niitä mahdollisuuksia, joita laskennallisten menetelmien käyttö yhdessä suurten aineistojen kanssa avaa eksploratiiviselle tutkimukselle. Artikkelin temaattisena kontekstina toimii digitaalisen parlamenttitutkimuksen tutkimusala, joka on viimeisten vuosien aikana – ensisijaisesti tutkimuskäyttöön avattujen aineistojen määrän kasvun myötä – avannut uusia näkökulmia parlamenttipuheen tutkimuksessa.[1]

FinParl-aineisto tietovarantona

Turun yliopiston eduskuntatutkimuksen keskus, Aalto-yliopiston semanttisen laskennan tutkimusryhmä (SeCo) sekä Helsingin yliopiston HELDIG-yksikkö saivat vuonna 2019 Suomen Akatemian suunnatussa DIGIHUM-rahoitushaussa kolmivuotisen rahoituksen vuosille 2020–2022 ”Semanttiset parlamentit”-konsortiohankkeen toteuttamiseksi. Yksi hankkeen työpaketeista keskittyi muodostamaan eduskunnan täysistuntopöytäkirjoista koneluettavan aineistokorpuksen, johon sisältyisivät kaikki eduskunnan täysistuntokeskustelut vuodesta 1907 eteenpäin. Työpaketin primaarimateriaalina olivat eduskunnan omana työnään PDF-muotoiseksi digitoimat täysistuntopöytäkirjat vuosilta 1907–2014 sekä vuodesta 2015 sähkösyntyisenä tarjolla olevat täysistuntopöytäkirjat. Pöytäkirjat ovat, avoimuus- ja julkisuuslain mukaisesti, tarjolla myös eduskunnan www-sivujen kautta, jolla sivustolla on myös lukuisia muita digitoituja tai sähköisiä, eduskunnan toimintaan liittyviä asiakirjoja ja tietolähteitä. Tutkimustyön kannalta avoimesti tarjolla olevat sähköiset aineistot ovat jo sinällään merkittävästi parantaneet aineistojen saatavuutta ja käytettävyyttä, kun fyysinen välimatka tutkijan ja eduskunnan toimitilojen välillä ei enää muodosta estettä aineistojen käytölle.

Vaikka PDF-muotoisena käytettävissä olevat täysistuntopöytäkirjat kattavat koko nykyisen yksikamarisen eduskunnan toimikauden vuodesta 1907 alkaen, teknisestä näkökulmasta PDF-muoto palvelee lähinnä aineistojen faksimile-tyyppistä käyttöä ajasta ja paikasta riippumatta, kun digitoituja pöytäkirjoja voi tarkastella sähköisesti niiden alkuperäisen ulkoasun mukaisessa muodossa. Ennen kokonaan sähköiseen arkistointiin siirtymistä syntyneet paperiasiakirjat on digitoinnin lisäksi käsitelty optisen tekstintunnistuksen (*Optical Character Recognition*, OCR) avulla, eli asiakirjojen tekstisisältö on liitetty PDF-dokumenttiin niin kutsuttuna tekstitasona kuvamuotoisten alkuperäissivujen ”päälle”. Tämä on nähtävissä, kun asiakirjaa merkitsee hiirellä ”maalaamalla”, jolloin merkitty teksti tulee näkyviin ja se voidaan esimerkiksi kopioida jatkokäyttöä varten. Vaikka teknisesti OCR on nykyisin varsin helppo toteuttaa aivan perusohjelmistoilla, liittyy prosessiin kaksi ongelmaa. Ensimmäiseksi, alkuperäisen

asiakirjan laatu vaikuttaa voimakkaasti tunnistuksen laatuun; mitä vanhempi tai huonolaatuisempi paperidokumentti, sitä enemmän tunnistettuun tekstiin sisältyy virheitä ja puutteita, eivätkä peruskäyttäjälle suunnatut ohjelmistot tarjoa kovinkaan hyviä työkaluja OCR-tuloksen parantamiseksi. Toiseksi, tunnistettu teksti noudattaa alkuperäisen dokumentin ladontaa ja ulkoasua, jolloin esimerkiksi valtiopäiväasiakirjoissa laajasti käytetty palstoitus luo omat haasteensa jatkokäytölle. Myös tavutukset ja aiemmin lihavoinnin tilalla käytetty sanan harvennus tuottavat ongelmia ja vaikuttavat huomattavasti PDF-lukijoiden hakutoiminnon tuloksiin, kun esimerkiksi tavutuksen katkaisemat sanat eivät löydy kokosanahaulla.

Nämä aineistojen digitointiin liittyvät ongelmat eivät toki ole ylitsepääsemättömiä - eivätkä monesti edes kovinkaan merkityksellisiä - jos aineistoja käytetään paperiaineistojen tapaan lähiluennan lähdeaineistona. Eli tutkimuksissa, joissa digitointi toimii aineistojen saatavuuden parantajana, itse tutkimustyön nojautuessa tutkijan itsensä suorittamaan lähteiden lukemiseen. Mutta niiden merkitys muuttuu marginaalisesta olennaiseksi, jos aineistoa on tarkoitus jatkoanalysoida laskennallisia, digitaalisia tutkimusmenetelmiä soveltamalla. Eli tutkimuksissa, joissa aineistoa ”louhitaan” käyttämällä etäluentaan kehitettyjä algoritmiperustaisia työkaluja, jotka käsittelevät aineistoa kokonaisuutena pyrkien esimerkiksi tunnistamaan siitä toistuvia, usein piileviä rakenteita tai havainnoimaan aineiston eri osien välisiä yhteyksiä.

Laskennallisten menetelmien soveltaminen vaatii digitoinnin ohella aineiston siirtämistä *koneluettavaan* muotoon. Digitointi ja koneluettavuus ovat siis kaksi eri asiaa, vaikka yhä edelleen niiden käytössä on paljon sekavuutta myös tutkijakunnan keskuudessa. Digitointi siis tarkoittaa alun perin ei-sähköisenä luodun aineiston siirtämistä sähköiseen muotoon. Vaikka usein puhutaan asiakirjojen digitoinnista, digitointi voi kohdista myös muihin esineisiin kuten esimerkiksi maalauksiin, huonekaluihin tai patsaisiin. Olennaista tässä on, että digitoinnin jälkeen aineistoa voidaan tarkastella myös tietokoneen ruudulta käsin. Koneluettava aineisto puolestaan on muokattua dataa, jossa alkuperäinen kohde on muokattu sellaiseen muotoon, että sitä voidaan käsitellä tietokoneella automaattisesti eli ilman ihmisen tekemiä interventioita ja hyödyntämällä algoritmeja ja laskennallisia menetelmiä. Koneluettava data on aina myös strukturoitua eli sen kohteena oleva ilmiö on mallinnettu noudattaen tiettyä ontologista rakennetta siten, että ilmiöön liittyvät semanttiset merkitykset eivät katoa. Tekstidokumenttien kohdalla strukturointi tapahtuu useimmiten rakentamalla dokumenteista sanapohjainen datarakenne, mutta vastaavasti esimerkiksi esineitä voidaan mallintaa kuvaamalla esine erilaisilla datayksiköillä - mitat, värit, painot jne. - mahdollisimman tarkasti. Se, millainen datarakenne valitaan, riippuu luonnollisesti myös siitä, millaiseen käyttöön koneluettava aineisto on tarkoitettu.

FinParl-aineiston datarakenteeksi valittiin vastaavissa eurooppalaisissa digitointihankkeissa yleisesti sovellettua, standardisoitua Parla-CLARIN-datamallia[2], joka siis on kehitetty nimenomaisesti parlamenttikeskustelujen koneluettavaksi datarakenteeksi. Primaariaineiston siirtäminen koneluettavaan muotoon koostui kolmesta vaiheesta. Ensinnäkin digitoidut täysistuntopöytäkirjat käsiteltiin uudelleen OCR-työkaluilla, mihin työvaiheeseen sisältyi myös OCR-tulosten laadunvarmistus eli käytännössä OCR-tuloksissa ilmenneiden virheiden korjaaminen siten, ettei lopulliseen data-aineistoon jäänyt tilastollisesti merkitsevää virhettä. Toisessa vaiheessa kukin puheenvuoro koostettiin yhtenäiseksi tekstiksi muun muassa yhdistämällä tavutuksen katkaisemat sanojen osat. Viimeisessä, kolmannessa vaiheessa aineisto järjestettiin istuntokohtaiseksi dataksi ja lisättiin puheenvuoro- ja istuntokohtaista metadataa. Jokaiseen puheenvuoroon lisättiin tieto puheenvuoron pitäjästä ja puheenvuoron istuntokohtainen, juokseva

järjestysnumero. Istuntokohtaiseen metadataan kuuluvat muun muassa tieto siitä, minkä valtiopäivän monesko istunto on kyseessä, jokaisen paikalla olleen kansanedustajan nimi, tiedot mahdollisista rooleista (puhemies, varapuhemies, ministeri tms.) sekä eduskuntaryhmä sekä istuntopäivä ja istunnon alkamis- ja päättymisaika. FinParl-aineisto on strukturoitu valtiopäivittäin XML-muotoisina tiedostoina, mikä ratkaisu sekä pitää yksittäisten tiedostojen koon hallittavana että helpottaa aineistonhallintaa, kun aineistosta voi ottaa käyttöön vain omaan tutkimushankkeeseen liittyvät valtiopäivät.[3]

Vuoden 2023 alusta FinParl-aineistoa on mahdollista käyttää *ParlamenttiSampo*-portaalin kautta, joka on loppukäyttäjille suunnattu, melko helppokäyttöinen datankäyttöportaali. *ParlamenttiSampo*-portaali rakentuu Aalto-yliopiston SeCo-tutkimusryhmän kehittämän Sampo-konseptin varaan, jossa käyttöön otettujen uusien tietomallien ja formaattien avulla data voidaan tarjota aiempaa rikkaammissa ja käyttökelpoisemmissa muodoissa tutkijoita ja sovellusten kehittäjiä varten. *ParlamenttiSampo*-portaalin avulla tietoa voi hakea ja selata tutkijoiden ohella myös laajempi yleisö ilman ohjelmointitaitoa. Portaali tarjoaa joukon erillisiä, mutta toisiinsa linkittyviä sovellusnäkymiä, joiden kautta palvelun sisältämää dataa voidaan hakea ja selata semanttista fasettihakua hyödyntäen temaattisesti.[4]

FinParl-aineisto ja sen mahdollistanut SEMPARL-konsortiohanke perustuvat linkitetyn datan konseptille, jossa olemassa olevia tietoaineistoja linkitetään toisiinsa tavoitteena rikastaa aineistoja samalla tiedon monistumista välttäen. Yhtenä esimerkkinä on FinParl-aineistoon sisältyvien kansanedustajien linkittäminen kansanedustajien biografia-aineistoihin, mikä tarjoaa loppukäyttäjälle mahdollisuuden hyödyntää koko rikastettua tietovarantoa omissa analyyseissään. Koska linkitetyssä datamallissa linkitetään tietovarantoja keskenään ja reaaliaikaisesti, näkyvät yhdessä tietovarannossa tehdyt muutokset suoraan linkitetyn datan tuloksissa.[5] Tämä vähentää oleellisesti virhelähteiden määrää ja siten osaltaan parantaa analyysitulosten luotettavuutta. Toki tässäkin pitää muistaa, että linkitetyn datan malli ei ole immuuni itse datassa oleville virheille, minkä lisäksi erityisesti automaattisiin linkitysmenetelmiin sisältyy aina riski virheellisten linkkien syntymisestä, jonka vaikutus riippuu vahvasti käytetystä linkitysalgoritmista. Esimerkkinä voidaan ajatella nimistöperustaiseen linkitykseen sisältyvää virheriskiä tilanteessa, jossa henkilön sukunimi on identtinen paikan, organisaation tai vastaavan kanssa. Mikäli viittauksen konteksti ei auta selvittämään viittauksen oikeaa kohdetta, voi linkitys tuottaa virheellisen viittauksen. Kun suurten aineistojen kohdalla jokaisen yksittäisen tietoentiteetin varmistaminen ei ole mahdollista, joudutaan tämän tyyppisissä järjestelmissä aina hyväksymään tietty epätarkkuus ja virheellisen tiedon osuus. Näitä voidaan kuitenkin merkittävästi pienentää kiinnittämällä kaikissa koneluettavaan muotoon siirtämisen prosessin vaiheissa riittävästi huomiota datan laadunvarmistukseen.[6]

Laskennallisten menetelmien tulo ihmistieteisiin on osaltaan kiihdyttänyt uudelleen väittelyitä laadullisen ja määrällisen tutkimuksen paremmuudesta, etenkin kun monet digitaalisista työkaluista esitellään laadullisen data-analyysin (*qualitative data-analysis*, QDA) välineinä.[7] Vaikka en itsekään pidä tiukkaa erottelua määrällisen ja laadullisen tutkimuksen välillä kovinkaan tarkoituksenmukaisena, jonkinlaiseksi muotitermiksi noussut *mixed method* ei mielestäni myöskään selvennä asiaa. Tutkimusmenetelmät ovat olennaisen tärkeä osa tutkimustyötä, mutta samalla niiden valinta ja käyttö on aina sidoksissa sekä aineistoon että tutkimuskysymyksiin. Kyse on siis tarkoituksenmukaisuusvalinnasta, jossa keskeistä on löytää välineet ekstrahoida eli erottaa aineistosta tietoa, joka auttaa vastaamaan esitettyihin kysymyksiin. Toisin sanoen, isoja aineistoja voidaan tutkia yhtä hyvin laadullisin kuin määrällisinkin menetelmin, joskin valtavan suurten aineistojen kohdalla laadullisia menetelmiä soveltava

tutkija joutuu väistämättä ratkaisemaan kysymyksen siitä, miten aineistosta lohkaistaan lähiluennalla hallittavissa oleva, luotettava ja edustava analyysiaineisto. Tässä kuitenkin laskennalliset menetelmät toimivat vain aineiston rajaamisen välineinä, eivät itse analyysimenetelmänä. Jos taas tuloksia haetaan laskennallisin menetelmin eli etäluennan kautta, tulee hyväksyä se, että tulokset muodostuvat kumulatiivisesti ja/tai aggregaatteina eli yksittäisten havaintojen kasautumana, jolloin etäluennan tulokset kuvaavat ja tyypittelevät aineistoa kokonaisuutena. Etäluennan tuloksia ei siten ole aina mahdollista - eikä usein edes mielekästä - palauttaa spesifiin, yksittäiseen aineiston elementtiin, koska laskennallisin menetelmin etsitään toistuvia malleja ja piileviä rakenteellisia piirteitä, joita voidaan parhaiten havainnoida lintuperspektiivistä. Laskennalliset menetelmät siis kasvattavat tarkasteluetäisyyttä, tavoitteena avata aineistoon uusia, lähiluennalle mahdottomia näkökulmia. Juuri edellä kuvatuista syistä pidän lähi- ja etäluentaa toinen toistaan täydentävinä menetelminä, joiden tarkoituksenmukainen käyttö mahdollistaa aineistojen moniulotteisen hyödyntämisen.

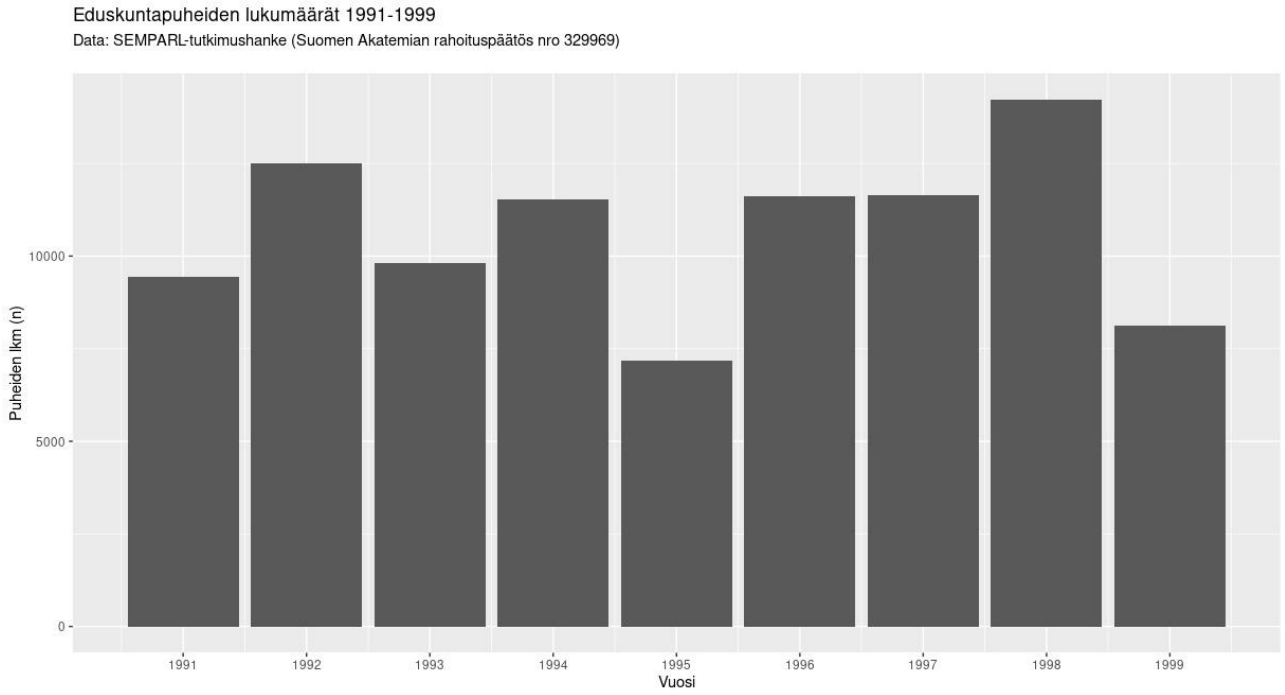
Digitoitujen, osin koneluettavassa muodossa tarjolla olevien parlamenttiaineistojen määrän kasvu ja siihen liittyvä laskennallisten tutkimuksen potentiaali on tunnustettu myös parlamenttipuheen ja poliittisten diskurssien tutkimuksen piirissä. Kuten historioitsija Deborah Kilroyn (2021) tuore analyysi, jossa hän tarkastelee puhujien sosiaalisten taustamuuttujien ja biografiadatan valossa 1600-luvun alun Englannin parlamentissa (*Journal of House of Commons*) pidettyjä puheita, osoittaa, digitoiduilla aineistoilla voidaan kaivautua hyvinkin kauas historiaan. Artikkelissa esitetty analyysi ei pelkästään täydennä kuvaamme tuon ajan parlamenttipuheen ja puhujien statusten välisestä suhteesta, vaan osoittaa erinomaisella tavalla, millaisia mahdollisuuksia suuret data-aineistot avaavat juuri eksploratiiviselle, tietokoneavusteiselle analyysille. Kiinnostava on myös Zoltan Majdikin[8] Yhdysvaltojen kongressin ilmastopuhetta analysoiva artikkeli, jossa kirjoittaja tutkii kongressin puheaineistoja semanttisten kontekstien sekä käytettyjen retoristen ilmausten näkökulmista. Maininnan arvoisina pidän myös Anamaria Dutceac Segestenin ja Michael Bossettan[9] sekä Thomas Jacobsin ja Robin Tschötschelin[10] määrällisiä ja laadullisia menetelmiä soveltavia artikkeleita, jotka molemmat hyödyntävät suuria poliittiseen puheeseen liittyviä aineistoja tarkastellen näitä sekä tekstinlouhinnan että perinteisen diskurssianalyysin keinoin. Omissa viimeaikaisissa tutkimuksissani olen hyödyntänyt laskennallisia menetelmiä esimerkiksi saksankielisten maiden valtionpäämiesten ja -naisten uudenvuodenpuheiden[11] sekä Suomen eduskunnassa käytyjen ympäristö- ja Eurooppa-keskustelujen[12] analysoinnissa.

Kansanedustajien eduskuntapuhe 1991–1999

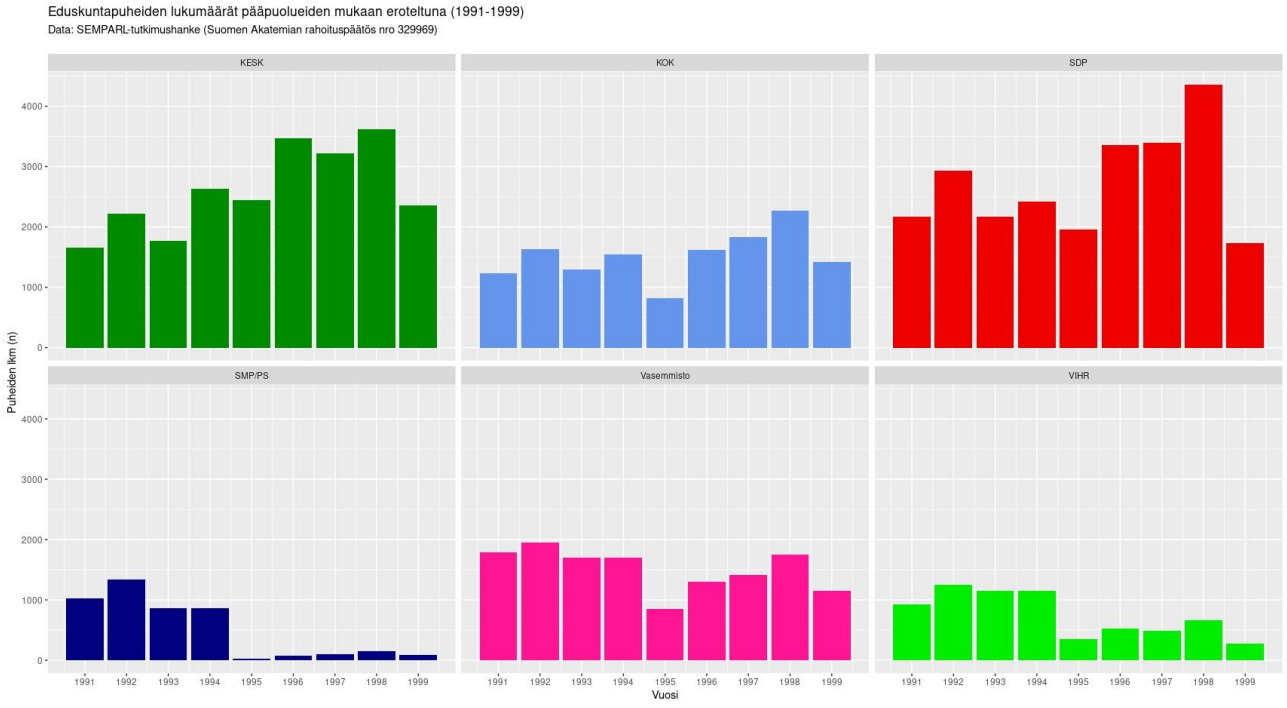
Havainnollistaakseni FinParl-aineiston mahdollisuuksia tutkimuskäytössä esittelen seuraavassa pienen koeporauksen kansanedustajien eduskuntapuheeseen vuosina 1991–1999. Koeporaukset tehdään politiikkatieteiden näkökulmasta, ja tavoitteena on analysoida, millä tavoin kansanedustajien keskinäiset suhteet rakentuvat, jos lähtökohdaksi otetaan kunkin edustajan semanttinen habitus, joka muodostetaan huomioimalla kansanedustajan tarkastelujaksolla eniten käyttämät käsitteet. Tämä näkökulma perustuu aiemmissa tutkimuksissa tehtyyn havaintoon, jonka mukaan kansanedustajien puhe heijastelee paitsi heidän edustamansa puolueen ideologista positiota, myös puolueen sijoittumista hallitus-oppositio-akselilla.[13] Jaan tarkastelun kahteen erilliseen aikaikkunaan – 1991–1994 ja 1995–1999 – jonka jaottelun perustana on Suomen liittyminen Euroopan unionin (EU) jäseneksi vuoden 1995 alusta. Näin pyritään tavoittamaan myös se parlamenttipuheen käsitteellinen ja/tai temaattinen muutos, joka mahdollisesti seurasi Suomen EU-jäsenyydestä. Analyysi itsessään rakentuu kolmessa vaiheessa. Aluksi kuvaan

analyysidatan valmistelun ja esitän datan pohjalta muutaman kuvailevan trendihavainnon. Tämän jälkeen testaan oletusta puoluetaustan vaikutuksesta tarkastelemalla sanastoeroja hallitus-oppositio-jaottelun perusteella. Analyysin viimeisessä osassa tarkennan yksittäisten kansanedustajien semanttiseen habitukseen ja teen sen perusteella havaintoja kansanedustajien semanttisesta läheisyydestä eli siitä, miten lähellä tai etäällä eri kansanedustajat toisistaan ovat ja mikä tätä läheisyyden astetta voisi selittää.

Analyysissä käytetään koneluettavasta FinParl-aineistosta louhittua dataa, joka saatiin erottamalla koko aineistosta vuosille 1991–1999 sijoittuvat puheenvuorot. Kaikki aineiston analyysit toteutin R-tilasto-ohjelmalla (https://www.r-project.org/), joka tarjoaa erinomaiset työkalut eksploratiiviselle data-analyysille ja laskennalliselle tekstianalyysille. Käytettävään analyysiaineistoon sisältyy yhteensä n. 96 000 yksittäistä puheenvuoroa. Puheenvuoroista on poistettu puhemiehistön pitämiksi merkityt puheenvuorot. Aineistossa on kyse OCR-käsitellystä osasta FinParl-aineistoa ja puheenvuoron pitäjien tiedoissa on jonkin verran virheellisyyttä. Puheenvuorojen pitäjien tiedoista 370 puhujaa on tunnistettu oikein ja 70 puhujan tiedot ovat tavalla tai toisella virheellisiä, eli puhujien nimistä 15 prosenttia olisi tässä otoksessa virheellisiä. Kuitenkin virheellisen puhujatiedon sisältävien puheenvuorojen osuus kaikista tarkastelujakson puheenvuoroista on vain 0,35 prosenttia (n=335), mikä on niin pieni osuus, ettei sillä ole vaikutusta tulosten luotettavuuteen. Datan valmisteluvaiheessa aineisto myös lemmatisoitiin, eli sanat palautettiin perusmuotoisiksi ja rikastettiin morfologisella tiedolla kunkin sanan sanaluokasta (niin kutsuttu part-of-speech- eli POS-tag). Lemmatisoinnin suoritin Turun yliopiston kieliteknologiaan erikoistuneen TurkuNLP-tutkimusryhmän kehittämän parserin avulla.[14] Lemmatisoinnin tuloksena muodostettu data sisälsi yhteensä noin 2,9 miljoonaa sanaa ja noin 490 000 uniikkia termiä. Myöhemmin tässä luvussa esiteltäviä sanastoanalyysejä varten redusoin tätä sanadataa poistamalla siitä kaikki niin kutsutut stopwordit eli semanttisesti merkityksettömät sanat, kuten konjunktiot sekä numerot, minkä jälkeen suodatin aineistosta pois kaikki muut sanaluokat kuin substantiivit, verbit, adjektiivit ja henkilösanat. Redusoituun aineistoon sisältyy hieman yli 2 miljoonaa sanaa ja vajaat 75 000 uniikkia termiä. Erityisesti jälkimmäisen ryhmän kohdalla havaittava suuri ero suodattamattoman ja suodatetun aineiston välillä kertonee siitä, että eduskuntapuheessa käytetään myös runsaasti erilaisia täytesanoja. Näiden tutkimus olisi oma, enemmän kielitieteellinen projektinsa, joka ei kuitenkaan sovi tämän artikkelin analyysin puitteissa toteutettavaksi. Edellä kuvaamani datan valmistelu ja redusointi noudattavat tekstintutkimuksessa yleisesti sovellettuja periaatteita.



Kuva 1a: Eduskunnan täysistuntopuheenvuorojen intensiteetin vaihtelu 1991–1999 kokonaisuutena.

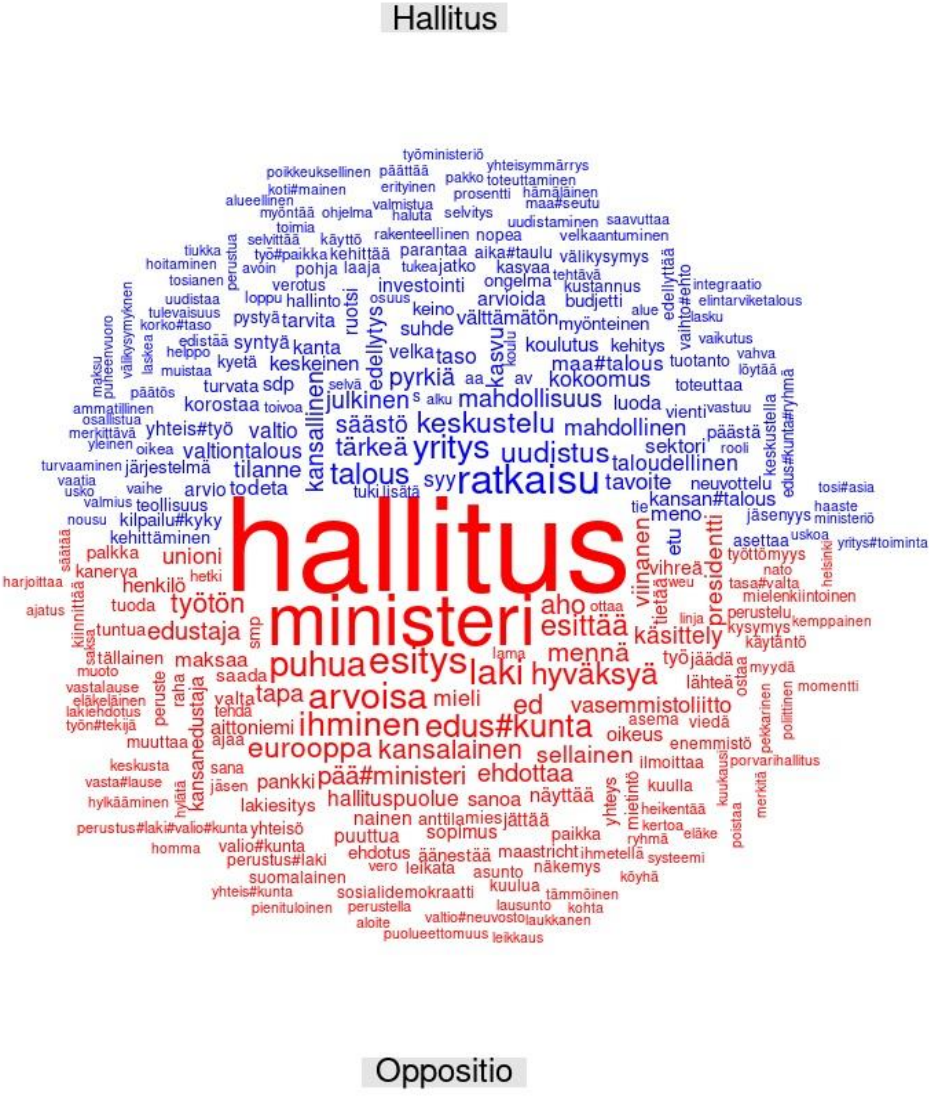


Kuva 1b: Eduskunnan täysistuntopuheenvuorojen intensiteetin vaihtelu 1991–1999 pääpuolueittain.

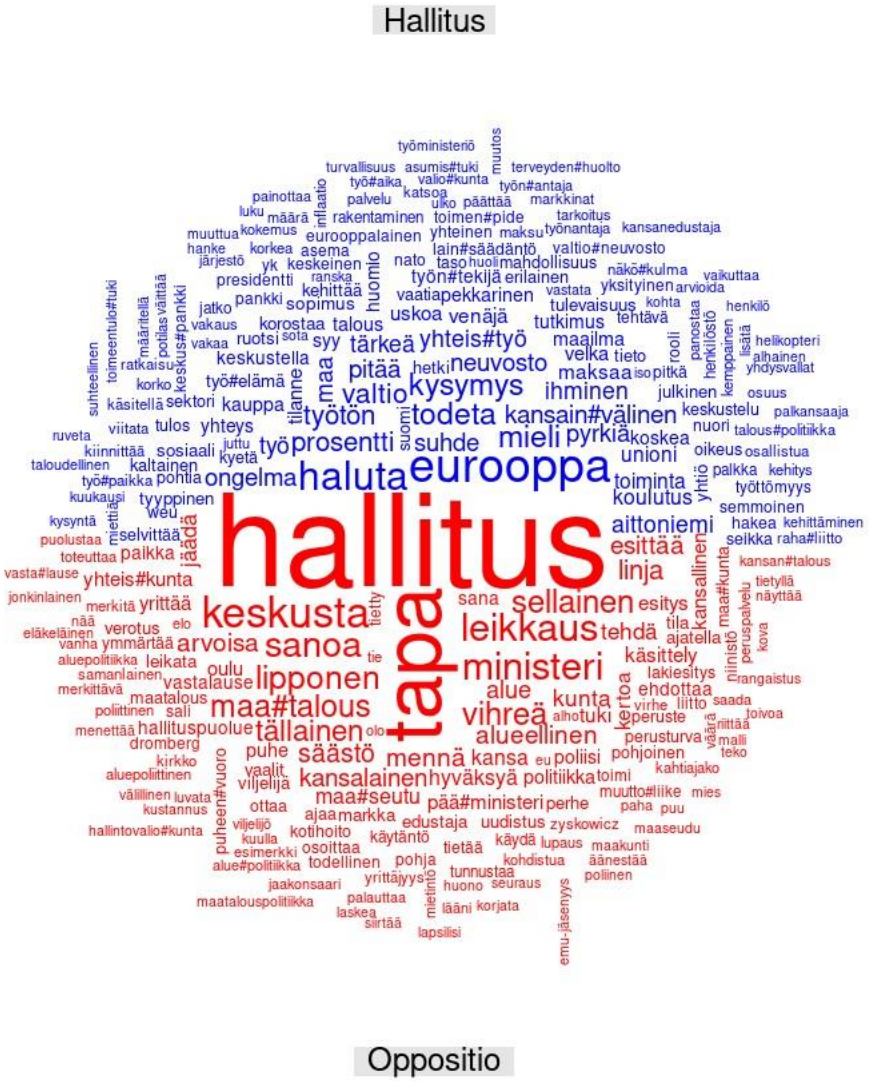
Kuva 1 esittää eduskunnan täysistunnoissa pidettyjen puheenvuorojen intensiteetin vaihtelun vuosina 1991 kokonaisuutena (diagrammi (a)) ja pääpuolueittain eriteltynä (diagrammi (b)). Analyysissä huomioin vain suurimmat eduskuntaryhmät, minkä lisäksi niputin tarkastelujaksolla eduskunnassa edustettuina olleet vasemmistopuolueet (SKDL, DEVA, SKP, Vasemmistoliitto) yhdeksi vasemmisto-ryhmäksi. Perustelen tätä ensisijaisesti analyysin pyrkimyksellä tuoda esiin eroja ja samankaltaisuuksia

ideologisesti erilaisten poliittisten ryhmien välillä. Vasemmisto-ryhmän alle kootut puolueet ovat ideologisesti hyvin samankaltaisia, minkä oletan myös heijastuvan sanastollisesti samankaltaisena puhuntana.

Eduskunnan täysistuntopuheenvuorojen määrä on pysynyt tarkastelujaksolla varsin vakaana: keskimäärin tarkastelujaksolla pidettiin noin 10 700 puheenvuoroa vuodessa. Selkeimmät poikkeukset ovat vuodet 1995 ja 1999, mitä selittänee eduskuntavaalien vuoksi lyhyemmäksi jäänyt istuntokausi. Kiinnostavaa on, että myös eduskunnassa edustettuina olleiden puolueiden käyttämien täysistuntopuheenvuorojen intensiteetti noudattaa varsin yhdenmukaisesti yleistä trendiä. Toisin sanoen, mikään yksittäinen puolue ei näyttäisi olevan yli- tai aliedustettuna, etenkin kun puheiden määrät suhteutetaan eduskuntaryhmien kokoon. Tämä näyttäisi tukevan ajatusta siitä, että puolueet näkevät täysistuntopuheen olennaisena osana edustuksellisen demokratian julkista toimintaa, mistä syystä ne pyrkivät myös olemaan esillä täysistuntokeskusteluissa.



Kuva 2a. Hallitus- ja oppositiopuolueille tyypilliset sanat 1991–1994.



Kuva 2b: Hallitus- ja oppositiopuolueille tyypilliset sanat 1995–1999.

Puhujakohtaisen sanaston tarkastelu sen valossa, kuuluuko puhuja puheen pitohetkellä hallitukseen vai oppositioon, osoittaa tämän valta-asetelman varsin kiistattoman vaikutuksen sanavalintoihin (Kuva 2). Valittu esitystapa, vertailusanapilvi, jossa on esitettynä 250 tarkastelujakson yleisimmin käytettyä sanaa, eroaa perinteisestä, sanojen frekvenssiin eli käyttömäärään perustuvasta sanapilvikuvaajasta siinä, että vertailusanapilvessä on mukana vähintään kaksi luokkaa (tässä esimerkissä siis ”hallitus” ja ”oppositio”). Vertailusanapilvi sijoittaa kunkin sanan sen luokan yhteyteen, jonka kohdalla sana käyttö poikkeaa tilastollisesti merkitsevästi eniten ylöspäin sanan keskimääräisestä käytöstä. Näin on mahdollista tunnistaa kullekin luokalle tyypilliset sanat, jotka erottavat ko. ryhmän kielenkäytön muiden ryhmien vastaavasta. Kyse on siis nimenomaisesti *erottelevista* sanoista, kaikkien ryhmien samankaltaisesti käyttämät sanat eivät näy sanapilvessä. Eri sanojen esittämiseen käytetty fonttikoko puolestaan korreloi sanan esiintymisfrekvenssin kanssa kunkin luokan sisällä. Näin sanapilvestä on myös

mahdollista paitsi tunnistaa tietyn luokan piirissä eniten käytetyt sanat, myös havainnoida saman luokan sisäistä varianssia sanojen käytössä: mitä vähemmän vaihtelua sanojen esittämiskoossa saman luokan sisällä on, sitä tasaisemmin sanoja on luokan piirissä käytetty, ja päinvastoin. On kuitenkin tärkeätä pitää mielessä, että fonttikokoon liittyvät vertailut toimivat vain luokan sisällä, eivätkä siten mahdollista vertailuja sanafrekvenssien osalta luokkien välillä.

Käytännössä sanapilvianalyysi vahvistaa tulkintaa, jonka mukaan oppositiopuolueiden tapa viitata puheenvuoroissaan hallitukseen kollektiivisesti tai yksittäisiin ministereihin jatkuu riippumatta siitä, mitkä puolueet kulloinkin ovat hallitusvastuussa, mitkä puolestaan oppositiossa. Tässä kohdin analyysi näyttäisi tuovan selvästi esille eduskuntapuheeseen liittyvän rakenteellisen konvention eli vakiintuneen tavan puhua, jonka mukaan juuri oppositiopuolueet korostavat puhunnassaan toimeenpanovaltaa käyttävän hallituksen toimintaa, ei vain tehden näin selväksi viestintänsä kohteen, vaan myös alleviivatakseen (usein kritiikin) kohteena olevasta asiasta vastuussa olevaa toimijaa. Näiden konventiotermien takaa löytyy kuitenkin hallituksen ja opposition temaattinen kamppailu, jota vuosina 1991–1994 näyttäisi leimanneen lamasta toipumisen ohella myös Suomen EU-jäsenyysratkaisua edeltänyt kiivas keskustelu. Lamaan liittyvinä käsitteinä nostan esimerkkeinä esille hallituksen osalta sellaiset termit kuin ”valtiontalous”, ”kasvu”, ”uudistus”, ”julkinen”, ”säästö”, ”kilpailukyky”, ”velka” ja ”koulutus”, opposition osalta puolestaan ”työtön”, ”kansalainen”, ”viinanen” (viittauksena valtiovarainministeri Iiro Viinaseen), ”yhteiskunta”, ”köyhä” ja ”pankki”. Kiintoisaa tässä on, että hallituspuolueiden sanastossa on runsaasti liike- ja toimintakäsitteiksi määriteltäviä sanoja, jotka korostavat tavoitteita ja toimintaa. Opposition kohdalla vastaavia käsitteitä on vähemmän, sanaston keskittyessä enemmänkin kuvailevaan käsitteistöön. EU-jäsenkysymyksen kohdalla mielenkiintoisena voi pitää sitä, että termit ”eurooppa”, ”unioni”, ”maastricht”, ”sopimus” ja ”yhteisö” esiintyvät useammin oppositiopuolueiden puheenvuoroissa, kun taas hallituksen puhunnassa selkeimmin EU-ratkaisuun liittyviksi voidaan katsoa termit ”integraatio”, ”maatalous”, ”keskustelu” ja ”mahdollisuus”. Näissäkin näyttäisi olevan jonkinasteinen painottuminen mahdollisuuksia ja tavoitteita korostavaan puhuntaan.

Vertailu vaalikauteen 1995–1999 on sikälikin mielenkiintoinen, että vuoden 1995 eduskuntavaaleissa pääministeripuolue vaihtui keskustasta SDP:hen, Paavo Lipposen noustessa Esko Ahon seuraajaksi. Avoimen EU-myönteinen Lipponen sai vastuulleen Suomen EU-jäsenyysajan ensimmäiset vuodet ja oli siten merkittävästi muokkaamassa Suomen EU- ja Eurooppa-politiikan linjaa. Kiinnostavaksi tämän aikakauden tekee myös se, että suurista puolueista kokoomus jatkoi hallituksessa myös Lipposen aikana.

Vuoden 1995–1999 sanapilvessä (Kuva 2b) voidaan havaita sama parlamenttipuheen konventio kuin vuosina 1991–1994 eli ”hallitus”-termin käytön painottuminen oppositiopuolueiden täysistuntopuheisiin. Tämä vahvistaa näkemystä siitä, että kyse on juuri konventiosta, parlamenttipuheen vakiintuneesta muotopiirteestä. Toinen kiinnostava havainto on mielestäni se, että termit ”maatalous”, ”maaseutu” ja ”alueellinen” siirtyvät opposition sanapilveen. Maatalouteen liittyvät teemat ovat perinteisesti olleet keskustan ydinteemoja, eli näiden termien siirtymistä selittäisi ensisijaisesti keskustan päätyminen oppositioon vuoden 1995 vaalien jälkeen. Tukea ajatukselle siitä, että Eurooppa-teema oli erityisesti aiemmin oppositiossa olleen SDP:n agendalla antaa osaltaan käsitteiden ”eurooppa”, ”unioni”, ”neuvosto”, ”yhteistyö” ja ”kansainvälinen” painottuminen hallituspuolueiden puhunnassa SDP:n noustua pääministeripuolueeksi. Tässä suhteessa on ilmeinen

jatkumo yli vaalikausien, samoin kuin siinä, että hallituksen puhunnassa korostuvat oppositiopuolueita enemmän tekemistä ja tavoittelemista korostavat liikekäsitteet kuten ”pyrkiä”, ”haluta”, ”uskoa”, ”vaatia” ja ”tulevaisuus”.

Edellä tehdyt havainnot perustuvat siis aineiston tarkasteluun lintuperspektiivistä, eniten käytettyjä sanoja painottaen. Tällainen makrotason tarkastelu auttaa näkemään pidemmän aikavälin sanaston rakennetta, joka puolestaan heijastelee ko. aikavälillä käsiteltyjä teemoja. Analyysiä voitaisiin jatkaa tämän jälkeen esimerkiksi KWIC- eli keyword-in-context -analyysillä, jossa tietokoneavusteisesti mallinnettaisiin eri avaintermien – esimerkiksi ”eurooppa” tai ”työtön” – ympärillä esiintyvien sanojen rakennetta, mikä kontekstinäkökulma auttaisi pohtimaan keskustelussa käsiteltyjä aiheita. Tästä voitaisiin edelleen edetä lähiluentaan saakka poimimalla KWIC-analyysin tulosten pohjalta alkuperäiset puheenvuorot tutkimusongelman kannalta merkityksellisistä aiheista. Koska tämän katsausartikkelin ensisijaisena tavoitteena on esitellä aineiston tarjoamia lähtökohtia, tyydyn tässä kohdin vain mainitsemaan tämän jatkoanalyysivaihtoehdon.[15]

Edellä esitetty analyysi antaa varsin selviä viitteitä siitä, että sekä puolueen oma ohjelmallinen agenda että rooli hallitus-oppositio-asemassa vaikuttavat puolueen edustajien kielellisiin valintoihin täysistuntopuheenvuoroissa. Kummankin tekijän vaikutuksista on saatu näyttöä jo aiemmissa tutkimuksissa [16], mutta seuraavassa tarkastelussa fokus on sen analysoinnissa, miten yhteyksiä kansanedustajien välille voidaan rakentaa analysoimalla heidän sanastollista koheesiotaan eli sitä, miten samankaltaisesti eri edustajat puhuvat. Tätä havainnollistetaan seuraavilla kahdella esimerkillä, joissa kummassakin kunkin edustajan osalta huomioidaan hänen 150 yleisimmin käyttämäänsä sanaa ja näiden frekvenssit eli käyttömäärä. Yleisimpinä sanoina on huomioitu ainoastaan semanttisesti merkitykselliset verbit, substantiivit, adjektiivit sekä erisnimet. Ensimmäisessä esimerkkianalyysissä lasketaan kunkin edustajan osalta hänen puhuntansa samankaltaisuus suhteessa muihin edustajiin. Tämä, kieliteknologiassa kosiniläheisyytenä tunnettu mittari laskee käytettyjen sanojen perusteella kahden dokumentin välisen samankaltaisuusasteen skaalalla [0,1], jossa 0 tarkoittaa, ettei dokumenteilla ole minkäänlaista samankaltaisuutta, kun taas arvo 1 tarkoittaa, että dokumentit ovat identtiset. Laskennassa huomioidaan paitsi käytetyt sanat, myös näiden sanojen käytön suhteellinen osuus koko dokumentin sanastosta. Toinen esimerkkianalyysi havainnollistaa, miten parlamenttipuheen teemoja ja kansanedustajien sijoittumista näiden teemojen suhteen voidaan kartoittaa hyödyntämällä kansanedustajien täysistuntopuheissa käyttämiä sanoja.

Taulukko 1: Kansanedustajien täysistuntopuheiden kosiniläheisyys käytettyjen suomenkielisten sanojen perusteella (10 eniten ja vähiten samankaltaista edustajaparia).

1991–1994					
Eniten samankaltaiset (10)			*Vähiten samankaltaiset (10)*		
Heikki Haavisto (KESK)	Pekka Haavisto (VIHR)	0.927	Matti Puhakka (SDP)	Ritva Laurila (KOK)	0.0659

Hannu Suhonen (SMP)	Tina Mäkelä (SMP)	0.876	Mats Nyby (SDP)	Olli-Pekka Heinonen (KOK)	0.0656
Jukka Gustafsson (SDP)	Outi Ojala (VAS)	0.873	Kalle Röntynen (KESK)	Mats Nyby (SDP)	0.0641
Heidi Hautala (VIHR)	Pekka Haavisto (VIHR)	0.871	Matti Puhakka (SDP)	Osmo Polvinen (VAS)	0.0607
Heidi Hautala (VIHR)	Heikki Haavisto (KESK)	0.867	Matti Puhakka (SDP)	Pekka Kivelä (KOK)	0.0588
Marjatta Vehkaoja (SDP)	Outi Ojala (VAS)	0.867	Matti Puhakka (SDP)	Paavo Lipponen (SDP)	0.0583
Heli Astala (VAS)	Jukka Gustafsson (SDP)	0.862	Jussi Ranta (SDP)	Matti Puhakka (SDP)	0.0542
Tina Mäkelä (SMP)	Tuulikki Hämäläinen (SDP)	0.862	Martti Pura (KESK)	Matti Puhakka (SDP)	0.0386
Pertti Salolainen (KOK)	Sulo Aittoniemi (SMP)	0.862	Iivo Polvi (VAS)	Matti Puhakka (SDP)	0.0207
Antti Kalliomäki (SDP)	Tina Mäkelä (SMP)	0.860	Mats Nyby (SDP)	Matti Puhakka (SDP)	0.00408

1995–1999

Eniten samankaltaiset (10)			*Vähiten samankaltaiset (10)*		
Jukka Gustafsson (SDP)	Outi Ojala (VAS)	0.873	Jaakko Laakso (VAS)	Matti Puhakka (SDP)	0.0777
Marjatta Vehkaoja (SDP)	Outi Ojala (VAS)	0.867	Kari Urpilainen (SDP)	Matti Puhakka (SDP)	0.0764
Pertti Salolainen (KOK)	Sulo Aittoniemi (KESK)	0.862	Mats Nyby (SDP)	Riitta Uosukainen (KOK)	0.0762
Asko Apukka (VAS)	Marjatta Stenius-Kaukonen (VAS)	0.855	Iiro Viinanen (KOK)	Matti Puhakka (SDP)	0.0735
Arja Ojala (SDP)	Outi Ojala (VAS)	0.852	Esko-Juhani Tennilä (VAS)	Matti Puhakka (SDP)	0.0732
Ismo Seivästö (KD)	Toimi Kankaanniemi (KD)	0.852	Mats Nyby (SDP)	Olli-Pekka Heinonen (KOK)	0.0656
Erkki Tuomioja (SDP)	Esko Aho (KESK)	0.849	Matti Puhakka (SDP)	Paavo Lipponen (SDP)	0.0583
Antti Kalliomäki (SDP)	Reijo Laitinen (SDP)	0.848	Jussi Ranta (SDP)	Matti Puhakka (SDP)	0.0542

Antti Kalliomäki (SDP)	Erkki Tuomioja (SDP)	0.846	Iivo Polvi (VAS)	Matti Puhakka (SDP)	0.0207
Antti Kalliomäki (SDP)	Esko Helle (VAS)	0.846	Mats Nyby (SDP)	Matti Puhakka (SDP)	0.00408

Taulukossa 1 on esitetty kymmenen (10) eniten ja vähiten samankaltaista kansanedustajaparia heidän täysistuntopuheissaan käyttämiensä 150 yleisimmän sanan perusteella laskettuna. Tulosten perusteella kolme huomiota näyttäisi perustellulta. Ensinnäkin, eniten samankaltaisimpien edustajien kohdalla puoluetausta näyttäisi varsin voimakkaasti selittävän puheiden samankaltaisuutta. Useimmissa vertailupareissa puhujat kuuluvat joko samaan puolueeseen tai ovat puoluetaustaltaan lähellä toisiaan vasemmisto-oikeisto-akselilla. Toiseksi, SMP:n edustajien samankaltaisuutta sekä kokoomuksen että SDP:n edustajien kanssa näyttäisi selittävän populistinen retoriikka, jossa korostuu etenkin niin kutsuttujen etabloituneiden puolueiden haastaminen politiikan eri areenoilla. Julkisuusvaikutuksen osalta eduskunnan täysistuntokeskusteluissa rakennettu retorinen vastakkainasettelu saman teeman ympärille on varmasti ollut SMP:lle tehokas tapa rakentaa omaa poliittista imagoaan omassa kannattajakunnassaan. Ja kolmanneksi, hallitus-oppositio-asetelmalla ei näyttäisi olevan juurikaan vaikutusta läheisyysasteikon ääripäissä eli hallitus-oppositio-asetelma ei näyttäisi juurikaan polarisoivan keskustelua. Aikajaksolla 1991–1994 oppositiossa olleiden vihreiden kahden edustajan, Pekka Haaviston ja Heidi Hautalan, vahva samankaltaisuus Heikki Haaviston (kesk) kanssa selittyy EU-jäsenyysratkaisulla. Hautala ja Pekka Haavisto lukeutuivat molemmat puolueensa Eurooppa-entusiasteihin, Heikki Haaviston toimiessa jäsenyysneuvottelujen aikaan ulkoasiainministerinä.

Taulukko 2: Esimerkkejä puhujaparien jaetusta sanastosta.

Puhujapari	Aikajakso	Jaettu sanasto[a)]
Heikki Haavisto (KESK) – Pekka Haavisto (VIHR)	1991–1994	suomi, mieli, hallitus, saada, tehdä, arvoisa, eurooppa, kysymys, tällainen, edus#kunta, maa, kanta, suomalainen, haluta,sellainen, osa, tapa, venäjä, ministeri,puhua, nähdä, pitää, ottaa, keskustelu,mahdollisuus, yhteys, tilanne, yritys, sopimus,jäädä, sanoa, käydä, kansain#välinen, käyttää,ajatella, alue, mennä, kansalainen (n=38)
Marjatta Vehkaoja (SDP) – Outi Ojala (VAS)	1991–1994	hallitus, tehdä, saada, mieli, kunta, kysymys, haluta, esitys, osa, ottaa, suomi, tapa, työ, sanoa, pitää, todeta, puhua, ministeri, valtio, syy, sellainen, esittää, laki, kanta, sosiaali, tietää, tilanne, ihminen, yhteys, edus#kunta, markka, lapsi, tapahtua (n=33)

Pertti Salolainen (KOK) – Sulo Aittoniemi (KESK)	1995–1999	suomi, eurooppa, hallitus, tilanne, pitää, kysymys, haluta, suomalainen, saada, todeta, edus#kunta, osa, sellainen, mieli (n=14)
Erkki Tuomioja (SDP) – Esko Aho (KESK)	1995–1999	suomi, hallitus, edus#kunta, tehdä, sellainen, osa, kysymys, eurooppa, saada, tällainen, liittyä, talous, poliittinen, tapa, kanta, arvoisa, ratkaisu, ongelma, mahdollisuus, pitää, tilanne, ottaa, haluta, syy, tarvita, käyttää, keskustelu, mieli, valtio, päätös, muutos, peruste (n=32)

[a] Kummankin puhujan osalta huomioitu 50 eniten käytettyä sanaa, joista erotettu molempien käyttämät sanat.

Taulukossa 2 on esimerkinomaisesti avattu samankaltaisimpien puhujaparien sanastollista samankaltaisuutta. Valitut puhujaparit on poimittu satunnaisesti kymmenen samankaltaisimman puhujaparin joukosta (Taulukko 1). Kunkin puhujaparin kohdalla on listattuna ne sanat, jotka esiintyvät kummankin puhujan 50 yleisimmin käyttämän sanan joukossa. Taulukko osoittaa selvästi, että samankaltaisesti puhuvat edustajat eivät pelkästään käytä samoja sanoja, vaan he myös käyttävät näitä sanoja suhteellisesti hyvin samantyyppisesti. Jälkimmäisestä kertoo juuri se, että jaettujen sanojen osuus 50 yleisimmin käytetyn sanan joukossa on varsin korkea. Ainoa selkeänä poikkeuksena on puhujapari Pertti Salolainen ja Sulo Aittoniemi, joiden eniten käyttämistä sanoista vain 14 on samoja. Tämä, yhdessä korkean kosiniläheisyyden kanssa, näyttäisi viittaavan siihen, että nämä kaksi edustajaa puhuvat hyvin samankaltaisesti, mutta painottavat eri sanoja puheissaan eri tavoin.

Yhteisten sanojen tarkastelu vahvistaa puhtaan kosiniläheisyystarkastelun kohdalla tehtyä oletusta, että Pekka Haaviston ja Heikki Haaviston puheiden korkea samankaltaisuus liittyisi juuri ulko- ja Eurooppa-politiikkaan, kuuluvathan sellaiset sanat kuten ”Suomi”, ”Eurooppa”, ”Venäjä” ja ”kansainvälinen” molempien puhujien yleisimmin käyttämiin sanoihin. Marjatta Vehkaojan ja Outi Ojalan puhunta puolestaan näyttäisi painottuvan sosiaalisiin kysymyksiin, mikä vahvistaa tulkintaa siitä, että täysistuntopuheet heijastelevat myös puolueiden ohjelmallisia ja ideologisia painotuksia.

Kokonaisuutena arvioiden edellä esitetyt koeporausluoteiset analyysit eduskunnan täysistuntopuheesta osoittavat, millaisia uusia mahdollisuuksia FinParl-aineistokorpukseen ja laskennallisten tutkimusmenetelmien käyttöön liittyy parlamenttipuheen osalta. Etenkin eksploratiivisen data-analyysin kautta näyttäisi olevan mahdollista ei vain muodostaa aiempaa paremmin käsitystä parlamenttipuheen sisällöistä ja siinä tapahtuvista muutoksista, vaan myös tutkia pitkällä aikavälillä ilmeneviä rakenteita ja näiden toistuvuutta. Omana mielenkiintoisena osa-alueena pidän yksittäisten kansanedustajien kielenkäytön pohjalta rakentuvia makroanalyysejä, joiden kautta on mahdollista tutkia, miten eduskunnassa vallitsevat roolijaot ja puolueiden ohjelmalliset linjaukset tulevat näkyviksi täysistunnon puheenvuoroissa.

Keskustelu

Tässä katsausartikkelissa esittelin Suomen eduskunnan täysistuntopuheenvuorot vuodesta 1907 lähtien koneluettavassa muodossa tarjoavan FinParl-aineistokorpuksen avaamia mahdollisuuksia parlamenttipuheen tutkimiseksi laskennallisia tutkimusmenetelmiä hyödyntämällä. Aineiston ja menetelmien tarjoamia mahdollisuuksia havainnollistin muutamien analyyttisten koeporausten muodossa, joiden ensisijaisena tavoitteena oli antaa tietoa siitä, miten yhteiskuntatieteille ominaisia kysymyksiä voitaisiin tutkia suurten aineistojen ja laskennallisten menetelmien osalta. Tämä lisäksi pyrin antamaan konkreettisia suuntaviivoja sille, miten FinParl-aineistokorpuksesta louhittuja havaintoja voidaan jatkotutkia yhdistämällä etäluennan laskennallisia menetelmiä ja lähiluentaan perustuvia laadullisen tutkimuksen menetelmiä.

Esiteltyjen tulosten perusteella rohkenen todeta, että sovellettujen menetelmien avulla on mahdollista löytää aineistoista olennaista tietoa kiinnostuksen kohteena olevasta ilmiöstä, tässä siis täysistuntokeskusteluissa esille nousseista teemoista sekä kansanedustajien roolien ja poliittisen taustan vaikutuksista täysistuntopuheeseen. Empiiriseltä kannalta tulokset antoivat selkeitä viitteitä siitä, että puhujien asema hallitus-oppositio-asetelmassa selittää heidän sanavalintojaan, minkä lisäksi tein kiinnostavia havaintoja puolueiden ohjelmallisten linjausten näkymisestä täysistuntopuheissa. Siten jo esittämäni melko suppea analyysi vahvisti useampia aiemmissa tutkimuksissa tehtyjä havaintoja parlamenttipuheen sisältöjä ja sanavalintoja määrittävistä tekijöistä.

Toivon, että tässä paperissa esitetyt koeporaukset toimisivat muille ihmistieteilijäkollegoilleni kannustimena kiinnostua laskennallisten tutkimusmenetelmien soveltamisesta omissa tutkimuksissaan. Ainakin itselleni on selvää, että laskennalliset menetelmät eivät korvaa perinteisiä tutkimusmenetelmiä, vaan täydentävät niitä avaten samalla uusia mahdollisuuksia hypoteesien muotoilemiseksi, uusien aineistojen hyödyntämiseksi tai uusien näkökulmien löytämiseksi vanhoihin aineistoihin.

[1] Esim. Måns Magnusson, Richard Öhrvall, Katarina Barrling & David Mimno, "Voices from the Far Right: A Text Analysis of Swedish Parliamentary Debates," *SocArXiv (*2018), https://doi.org/10.31235/osf.io/jdsqc; Zoltan P. Majdik, "A Computational Approach to Assessing Rhetorical Effectiveness: Agentic Framing of Climate Change in the Congressional Record, 1994–2016," *Technical Communication Quarterly* 28, no. 3 (2019), 207–222, https://doi.org/10.1080/10572252.2019.1601774; Kilroy, Deborah, "All the king's men? A demographic study of Opinion in the first English Parliament of James I, 1604–10," *Parliaments, Estates and Representation* 41, no. 1 (2021), 1–23, https://doi.org/10.1080/02606755.2020.1857546; Jens Edlund, Daniel Brodén, Mats Fridlund, Cecilia Lindhé, Leif-Jöran Olsson, Magnus P. Ängsal & Patrik Öhberg, "A multimodal digital humanities study of terrorism in Swedish politics: An interdisciplinary mixed methods project on the configuration of terrorism in parliamentary debates, legislation, and policy networks 1968–2018," teoksessa *Intelligent Systems and Applications*, toim. Kohei Arai (Cham. Springer International Publishing, 2022), 435–449, https://doi.org/10.1007/978-3-030-82196-8_32; Kimmo Elo, "Debates on European integration in the Finnish parliament (Eduskunta)," teoksessa *Parliamentary Data in Action (DiPaDA 2022) Workshop*, toim. Matti La Mela, Fredrik Norén, Eero Hyvönen (CEUR Workshop Proceedings, vol. 3133, 2022), 129–145, http://ceur-ws.org/Vol-3133/paper09.pdf.

[2] https://clarin-eric.github.io/parla-clarin/ (linkki tarkistettu 25.10.2022).

[3] Koneluettavan FinParl-aineiston XML-rakennekuvaus, ks. Laura Sinikallio, *Eduskunnan täysistuntojen pöytäkirjojen muuntaminen semanttiseksi dataksi ja julkaiseminen verkkopalveluna*. Pro gradu -tutkielma, Aalto yliopisto. Sinikallio, 2022, 23–25, http://urn.fi/URN:NBN:fi:hulib-202204201707.

[4] Eero Hyvönen, Laura Sinikallio, Petri Leskinen, Senka Drobac, Jouni Tuominen, Kimmo Elo, Matti La Mela, Mikko Koho, Esko Ikkala, Minna Tamper, Rafael Leal & Joonas Kesäniemi, "Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet," *Informaatiotutkimus* 40, no. 3 (2021), 229–230. DOI: https://doi.org/10.23978/inf.107899.

[5] Ks. tark. Hyvönen, "Parlamenttisampo".

[6] Myös Kimmo Elo, "Big data, bad metadata a methodological note on the importance of good metadata in the age of digital history," teoksessa *Digital Histories: Emergent Approaches within the New Digital History*, toim. Mats Fridlund, Mila Oiva & Petri Paju (Helsinki: Helsinki University Press, Helsinki, 2020), 103–111, https://doi.org/10.33134/HUP-5-6.

[7] Gregor Wiedemann, "Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences," *Forum: Qualitative Social Research* 14, no. 2 (2013), 332–357, https://doi.org/10.17169/fqs-14.2.1949.

[8] Majdik, "Assessing Rhetorical Effectiveness".

[9] Anamaria Dutceac Segesten & Michael Bossetta, "Can Euroscepticism Contribute to a European Public Sphere? The Europeanization of Media Discourses on Euroscepticism across six Countries," *Journal of Common Market Studies*, 57. no. 5 (2019), 1051–1070, https://doi.org/10.1111/jcms.12871.

[10] Thomas Jacobs & Robin Tschötschel, "Topic models meet discourse analysis: a quantitative tool for a qualitative approach," *International Journal of Social Research Methodology*, 22, no. 5 (2019), 469–485, https://doi.org/10.1080/13645579.2019.1576317.

[11] Kimmo Elo, "A Text Network Analysis of Discursive Changes in German, Austrian and Swiss New Year's Speeches 2000-2021," *Digital Humanities Quarterly* 16, no 1 (2022), http://www.digitalhumanities.org/dhq/vol/16/1/000598/000598.html.

[12] Kimmo Elo & Jenni Karimäki, "Luonnonsuojelusta ilmastopolitiikkaan. Ympäristöpoliittisen puhunnan muutos eduskuntakeskusteluissa 1960–2020," *Politiikka* 63, no 4 (2021), 373–402, https://doi.org/10.37452/politiikka.109690; Elo, "Debates on European integration".

[13] Esim., Liesbet Hooghe, Gary Marks & Carole J. Wilson, "Does Left/Right Structure Party Positions on European Integration?," *Comparative Political Studies* 35, no 8 (2002), 965–989, https://doi.org/10.1177/0010414002236310; Achim Hurrelmann, Stephanie Kerr, Anna Gora & Philipp Eibl, "Framing the Eurozone crisis in national parliaments: is the economic cleavage really declining?," *Journal of European Integration* 42, no. 4 (2019), 489–507, https://doi.org/10.1080/07036337.2019.1658755; Frank Wendler, "Contesting Europe, or Germany's Place in Europe? European Integration and the EU Policies of the Grand Coalition Government in the Mirror of Parliamentary Debates in the Bundestag, *German Politics,* 20, no. 4 (2011), 486–505, https://doi.org/10.1080/09644008.2011.606314; Marie-Eve Bélanger & Frank Schimmelfennig, "Politicization and rebordering in EU enlargement: membership discourses in European parliaments," *Journal of European Public Policy*, 28, no. 3 (2021), 407–426, https://doi.org/10.1080/13501763.2021.1881584.

[14] Kyseessä on luonnollisen kielen automaattiseen, morfologiseen prosessointiin kehitetty neuroverkkopohjainen parseri, ks. tark. https://turkunlp.org/Turku-neural-parser-pipeline/ (linkki tarkistettu 7.10.2022).

[15] KWIC-analyysin toteuttamisesta, ks. esim., Matthew L. Jockers & Rosamond Thalken, *Text Analysis with R* (Heidelberg: Springer International Publishing, 2020), 99ff.

[16] Esim. Hooghe et al., "Left/Right Party Positions"; Magnusson et al., "Voices from the Far Right"; Bélanger & Schimmelfennig, "Politicization and rebordering in EU enlargement".