RAJAPINNOILLA-TEEMANUMERO

Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet

Eero Hyvönen[1,2]
eero.hyvonen@aalto.fi
https://orcid.org/0000-0003-1695-5840

Laura Sinikallio[1]
laura.sinikallio@helsinki.fi
https://orcid.org/0000-0001-7398-6585

Petri Leskinen[2]
petri.leskinen@aalto.fi
https://orcid.org/0000-0003-2327-6942

Senka Drobac[1,2]
senka.drobac@aalto.fi
https://orcid.org/0000-0002-7645-3079

Jouni Tuominen[2,1]
jouni.tuominen@aalto.fi
https://orcid.org/0000-0003-4789-5676

Kimmo Elo[3]
kimmo.elo@utu.fi
https://orcid.org/0000-0002-3223-5221

Matti La Mela[1]
matti.lamela@helsinki.fi
https://orcid.org/0000-0003-0340-9269

Mikko Koho[2]
mikko.koho@aalto.fi
https://orcid.org/0000-0002-7373-9338

Esko Ikkala[2]
esko.ikkala@aalto.fi
https://orcid.org/0000-0002-9571-7260

Minna Tamper[2]
minna.tamper@aalto.fi
https://orcid.org/0000-0002-3301-1705

Rafael Leal[2]
rafael.leal@aalto.fi
https://orcid.org/0000-0001-7266-2036

Joonas Kesäniemi[2]
joonas.kesaniemi@aalto.fi
https://orcid.org/0000-0002-3770-0006

1 Helsingin yliopisto, Digitaalisten ihmistieteiden keskus HELDIG, SeCo-tutkimusryhmä
2 Aalto-yliopisto, tietotekniikan laitos, SeCo-tutkimusryhmä
3 Turun yliopisto, Eduskuntatutkimuksen keskus

Semanttinen parlamentti -hankkeessa 2020–2022 luodaan eduskunnan tietokannoista ja niihin liittyvistä muista aineistoista uudenlainen linkitetyn avoimen datan (Linked Open Data, LOD) palvelu, tietoinfrastruktuuri ja semanttinen portaali *Parlamenttisampo – eduskunta semanttisessa webissä*, joiden avulla tutkitaan poliittista kulttuuria ja kieltä. Dataa linkittämällä voidaan rikastaa eduskuntadataa muilla tietolähteillä kuten biografisella tiedolla, terminologioilla ja lainsäädännön dokumenteilla. Parlamenttisampo on kieli- ja semanttisen webin teknologioihin perustuva palvelukokonaisuus tutkijoita, kansalaisia, mediaa ja valtionhallintoa varten. Artikkelissa esitellään hankkeen visio, ensimmäisiä tuloksia ja niiden hyödyntämismahdollisuuksia: eduskunnan kaikkien täysistuntojen 1907–2021 yli 900 000 puheesta on valmistunut linkitetyn datan tietämysgraafi (knowledge graph); data on myös saatavilla XML-muodossa, jossa hyödynnetään uutta kansainvälistä Parla-CLARIN-formaattia. Ensimmäistä kertaa eduskunnan puheiden koko aikasarja on muunnettu dataksi ja datapalveluksi yhtenäisessä muodossa. Lisäksi puheet on yhdistetty eduskunnan kansanedustajien tietokannasta luotuun ja muista tietolähteistä rikastettuun toiseen tietämysgraafiin laajemmaksi ontologiaperustaiseksi datapalveluksi Finn-Parla. Datapalvelua voidaan käyttää eduskuntatutkimukseen parlamentaarisesta ja edustuksellisesta kulttuurista sekä poliittisen kielen käytöstä analysoimalla kansanedustajien täysistunnoissa pitämiä puheita ja poliitikkojen verkostoja data-analyysin keinoin. Palvelun rajapinnan avulla voidaan myös kehittää eri käyttäjäryhmille sovelluksia, kuten hankkeessa valmistuva Parlamenttisampo-portaali.

Asiasanat: eduskuntatutkimus, datapalvelut, semanttinen web, linkitetty data

Pysyvä osoite: https://doi.org/10.23978/inf.107899

Johdanto

Semanttinen parlamentti -hanke[1] on osa Suomen Akatemian rahoittamaa DIGIHUM 2020–2022 -ohjelmaa[2] ja perustuu yhteistyöhön Helsingin yliopiston (Digitaalisten ihmistieteiden keskus HELDIG), Aalto-yliopiston (tietotekniikan laitos) ja Turun yliopiston (Eduskuntatutkimuksen keskus) välillä. Hankkeen innovaationa on tietämyksen muodostamisen (information/knowledge extraction) (Martínez-Rodríguez ym., 2020), semanttisen webin teknologioiden[3] (Hyvönen, 2018), linkitetyn datan julkaisuperiaatteiden (Heath & Bizer, 2011) ja yleiseurooppalaisen CLARIN-infrastruktuurin[4] Parla-CLARIN-formaatin[5] hyödyntäminen parlamentaaristen aineistojen julkaisemisessa verkossa ja käyttämisessä digitaalisten ihmistieteiden tutkimuksessa (McCarty, 2005; Gardiner & Musto, 2015). Sovellusalueena on monitieteinen eduskuntatutkimus (Benoît & Rozenberg, 2020). Hankkeessa kehitetään luonnollisen kielen käsittelyn (natural language processing) teknologiaa, jonka avulla puheista ja muista teksteistä voidaan muodostaa semanttisia, ts. tietokoneen "ymmärtämiä" rakenteita, dataa. Semanttista dataa hyödynnetään semanttisen webin teknologioilla, millä tavoitellaan monia etuja: 1) Linkitetty data ja ontologiat (Staab & Studer, 2009) tarjoavat viitekehyksen, jonka avulla voidaan harmonisoida heterogeenisiä hajautettuja tietoaineistoja ja yhdistää niitä toisiinsa sisällöllisesti laajemmiksi ja rikkaammiksi kokonaisuuksiksi. 2) Linkitetyn datan predikaattilogiikkaan perustuva semantiikka tarjoaa mahdollisuuden aineistojen rikastamiseen päättelemällä uutta tietoa. 3) Kun kone ymmärtää aineistojen sisällöllisiä merkityksiä, semantiikkaa, voidaan toteuttaa helpommin älykkäitä verkkopalveluita ja data-analyysejä. 4) Standardimuotoisen datan julkaisemiseen, käsittelyyn ja analyysiin voidaan käyttää uudelleen muiden toimijoiden toteuttamia valmiita työkaluja ja aineistoja; pyörää ei tarvitse keksiä uudelleen itse.

Datan julkaisussa ja käytössä sovelletaan Sampo-mallia (Hyvönen, 2021), jonka avulla kaikkien julkaisuprosessiin osallistuvien tahojen datalle voidaan luoda uutta lisäarvoa yhteistyön kautta ja samalla tarjota loppukäyttäjille aiempaa rikkaampia ja monipuolisempia verkkopalveluita ja dataa. Lähestymistapaa on sovellettu aiemmin "Sampo-sarjan" järjestelmissä[6], joilla on ollut

1 https://seco.cs.aalto.fi/projects/semparl/en/
2 https://www.aka.fi/tutkimusrahoitus/ohjelmat-ja-muut-rahoitusmuodot/akatemiaohjelmat/digitaaliset-ihmistieteet---digihum-2016-2022/
3 https://www.w3.org/standards/semanticweb/
4 https://www.clarin.eu/
5 https://clarin-eric.github.io/parla-clarin/
6 https://seco.cs.aalto.fi/applications/sampo/

sammosta riippuen parhaimmillaan miljoonia käyttäjiä verkossa[7]. Loppukäyttäjän kannalta sampo koostuu kahdesta pääkomponentista: 1) linkitetyn avoimen datan palvelusta ja 2) sen avulla kehitetystä portaalisovelluksesta, joka toimii esimerkkinä datapalvelun käytöstä sovellusten kehittämiseen. Semanttinen parlamentti -hankkeessa kehitettävä Parlamenttisampo on uusi jäsen 16 sammon sarjassa, joka perustuu n. 20 vuoden aikana asteittain kehittyneeseen avoimeen ja kollaboratiiviseen toimintamalliin ja sen työkaluihin kuten käyttöliittymien kehittämisessä käytettävään Sampo-UI-kehysjärjestelmään (Ikkala ym., 2021). Sampo-malli on sopusoinnussa modernien FAIR-periaatteiden[8] kanssa tavoitteena edistää datan löydettävyyttä (Findable), saavutettavuutta (Accessible), yhteentoimivuutta (Interoperable) ja uudelleenkäytettävyyttä (Reusable).

Semanttinen parlamentti -hankkeessa luodaan avoin kotimainen ontologia- ja tietoinfrastruktuuri eduskunnan aineistojen julkaisemiseksi ja rikastamiseksi linkitettynä avoimena datana (Linked Open Data, LOD). Työ yhdistyy ontologiainfrastruktuurin osalta laajempaan hankekokonaisuuteen Linked Open Data Infrastructure for Digital Humanities in Finland (LODI4DH)[9]. Sen tavoitteena on rakentaa kansallista digitaalisten ihmistieteiden linkitetyn datan tietoinfrastruktuuria osana Suomen Akatemian tutkimusinfrastruktuurien tiekarttaa ja FIN-CLARIAH-hanketta[10]. Uutta infrastruktuuria ja sille kehitettäviä työkaluja pilotoidaan eduskunnan historian, poliittisen kulttuurin ja kielen tutkimuksessa.

Tässä katsauksessa esitellään ensin tarkemmin visio eduskunnan aineistoista semanttisessa webissä, aiheeseen liittyvää aiempaa tutkimusta Suomessa ja maailmalla ja sitten Parlamenttisammon kehittämisessä saatuja ensimmäisiä konkreettisia tuloksia. Eduskunnan täysistuntojen pöytäkirjoista 1907–2021 on luotu tietämysgraafi[11] (knowledge graph) (Noy ym., 2019; Abu-Salih, 2020) ja siihen on linkitetty eduskunnan kansanedustajien tietokannasta muodostettu ja rikastettu tietämysgraafi. Näiden kokonaisuus on julkaistu linkitetyn datan palveluna Linked Data Finland -alustalla. Lopuksi tarkastellaan datapalvelun erilaisia käyttötapoja tutkimuksessa esimerkkien

7 Maamme yleisten kirjastojen ylläpitämää Kirjasampo.fi-palvelua, joka oli aluksi osa laajempaa Kulttuurisampo.fi-palvelua, käytti v. 2020 lähes kaksi miljoonaa asiakasta. Toiseksi käytetyin sampo on ollut kansainvälisesti palkittu, toisen maailmansodan aineistoihin perustuva Sotasampo.fi, jonka sivuilla on vieraillut yli 740 000 käyttäjää.

8 https://www.go-fair.org/fair-principles/

9 Tätä hanketta on esitelty tarkemmin sivulla https://seco.cs.aalto.fi/projects/lodi4dh/.

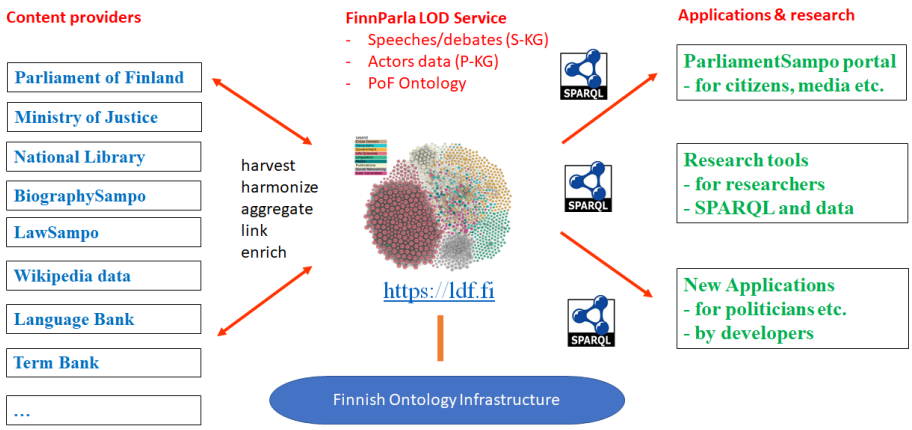
10 https://www.aka.fi/suomen-akatemian-toiminta/toimielimet/infrastruktuurikomitea/suomen-tutkimusinfrastruktuurien-tiekartta/

11 Tieto esitetään semanttisessa webissä verkkomuodossa eli graafeina.

avulla ja esitellään kehitteillä olevaa Parlamenttisampo-portaalia[12]. Lopuksi pohditaan Parlamenttisammon uusia käyttömahdollisuuksia yleisemmin eduskuntatutkimuksessa.

Visio: eduskunta semanttisessa webissä

Semanttinen parlamentti -hankkeen visiona on kehittää ja ottaa käyttöön living laboratory -ympäristössä kuvassa 1 esitetty julkaisumalli ja -prosessi eduskunnan aineistojen julkaisemiseksi ja hyödyntämiseksi datana semanttisessa webissä. Työssä keskitytään erityisesti kahteen ydinaineistoon kattaen eduskunnan koko historian vuodesta 1907 alkaen:



Kuva 1. Semanttinen Parlamentti -hankkeen linkitetyn datan julkaisumalli tiedon tuottajilta Parlamenttisampoon, digitaalisten ihmistieteiden tutkimuskäyttöön ja uusiin sovelluksiin.

1. Täysistuntojen puheenvuorot 1907–2021. Hanke julkaisee eduskunnan kaikkien täysistuntojen puheenvuorot vuodesta 1907 alkaen ensimmäistä kertaa yhtenäisessä muodossa 1) linkitettynä avoimena datana tietämysgraafina ja 2) Parla-CLARIN-formaatissa tutkimus- ja sovelluskäyttöön. Työn tuloksena syntynyt data, sen tuotantoprosessi ja datapalvelu on kuvattu tarkemmin julkaisussa (Sinikallio ym., 2021). Aiemmin aineistoja on ollut saatavilla, mutta vain osittain joko painettuina, PDF-muodossa, tekstinä, HTML-sivuina tai XML-muo-

12 Portaali tullaan julkaisemaan osoitteessa `https://parlamenttisampo.fi`.

dossa ajanjaksosta ja julkaisusta riippuen. Käytämme jatkossa tästä tietämysgraafista kuvan 1 mukaista nimitystä *S-KG* (Speech Knowledge Graph).

2. Kansanedustajien tietokanta 1907–2021. Täysistuntoaineistoihin liittyen julkaistaan kansanedustajien toimintaa 1907–2021 kuvaava uusi biografinen/prosopografinen[13] datapalvelu. Sen luomisprosessi ja rakenne on kuvattu tarkemmin julkaisussa (Leskinen ym., 2021). Tämän tietämysgraafin ytimenä on eduskunnan kansanedustajatietokanta, jota on rikastettu muista lähteistä, kuten Suomalaisen Kirjallisuuden Seuran Kansallisbiografiasta, valtioneuvoston tietokannalla hallituksista ja ministereistä, Biografiasammosta[14] ja Wikidatasta[15]. Käytämme jatkossa tästä graafista kuvan 1 mukaista nimitystä *P-KG* (Prosopographical Knowledge Graph).

Kuvassa 1 vasemmalla näkyy organisaatioita ja palveluita, jotka tuottavat omissa paikallisissa datasiiloissaan eduskunnan aineistoihin eri tavoin liittyvää tietoa, mutta keskenään epäyhteentoimivissa muodoissa. Dataa harmonisoidaan ja linkitetään esimerkiksi Lakisammon (LawSampo) aineistoihin, joka julkaisee Suomen lainsäädäntöä ja oikeustapauksia avoimena linkitetyn datan palveluna Semanttinen Finlex (Oksanen ym., 2019) ja loppukäyttäjille suunnattuna semanttisena portaalina Lakisampo[16] (Hyvönen ym., 2021a).

Semanttinen parlamentti -hankkeessa paikallinen data muunnetaan yhteentoimivaan (interoperable) muotoon ja julkaistaan uudessa FinnParla Linked Open Data -palvelussa (kuvassa 1 keskellä) Linked Data Finland -alustalla[17] (Hyvönen ym., 2014). Sen tietomallin perustana on 1) eduskunnan toimintaa kuvaava uusi ontologia (PoF Ontology) ja 2) joukko siihen liittyviä LODI4DH-ontologiainfrastruktuurin sanastoja ja ontologioita, joilla kuvataan esimerkiksi (historiallisia) paikkoja[18], ammatteja (Koho ym., 2019), henkilöitä ja organisaatioita. Data on hyödynnettävissä W3C:n piirissä luotujen linkitetyn datan standardien ja julkaisuperiaatteiden[19] mukaisesti (Heath & Bizer, 2011) datapalvelun rajapintojen kautta.

13 Biografiassa tutkitaan yksittäisten ihmisten elämää, kun taas prosopografiassa on kyse tietynlaisten ihmisjoukkojen, esimerkiksi eri ammattiryhmien tai puolueen jäsenten tutkimuksesta (Verboven ym., 2007).

14 https://seco.cs.aalto.fi/projects/biografiasampo/

15 https://www.wikidata.org/wiki/Wikidata:Main_Page

16 http://seco.cs.aalto.fi/projects/lawlod/

17 https://www.ldf.fi/

18 https://seco.cs.aalto.fi/projects/histoplaces/

19 https://www.w3.org/standards/semanticweb/data

Kuvan 1 oikeassa laidassa on kuvattu FinnParla-datapalvelun hyödyntämistapoja Parlamenttisampo-portaalina, rajapintojen ja työkalujen kautta suoraan tutkimustyössä ja data-analyyseissä sekä uusien sovellusten kehittämisessä. Keskeinen datan käyttötapa on SPARQL-kyselykielen[20] hyödyntäminen eduskuntatutkimuksen data-analyysejä varten, mutta dataa avataan ladattavaksi myös esimerkiksi CSV- (Comma Separate Values)[21] ja XML-muodoissa. Datapalvelun hyödyntämistä sovelluksissa demonstroidaan tekeillä olevassa, eri käyttäjäryhmille suunnatussa semanttisessa portaalissa *Parlamenttisampo – eduskunta semanttisessa webissä*. Portaalin käyttö ei edellytä SPARQL-kyselykielen hallintaa tai ohjelmointiosaamista. Toivon mukaan datapalvelua tullaan hyödyntämään myös tulevissa uusissa kolmansien osapuolien tutkimushankkeissa ja sovelluksissa. Parlamenttisammon datapalvelu oli koekäytössä keväällä 2021 pidetyssä Helsinki Digital Humanities Hackathonissa[22].

Parlamenttisammon datapalvelu ja siihen liittyvä portaali ovat tätä kirjoitettaessa hankkeen tutkijoiden sisäisessä tutkimuskäytössä, mutta ne julkaistaan hankkeen loppuun 31.12.2022 mennessä kaikille avoimesti hyödynnettäväksi CC BY 4.0-lisenssillä[23], joka mahdollistaa myös kaupallisen käytön.

Tutkimusta Suomessa ja maailmalla

Parlamenttiaineistoja käytetään tutkimuksessa monilla aloilla, sillä parlamenttien toiminta tuottaa monipuolista tietoa demokraattisten järjestelmien tilasta ja toiminnasta, poliittisesta elämästä ja yleisemmin kielestä sekä kulttuurista. Parlamenttien päätehtävät ovat uusien lakien säätäminen, hallituksen työn valvominen sekä valtion budjetista päättäminen. Näkyvin osa parlamenttien työtä ovat julkiset täysistunnot, joissa kansanedustajat keskustelevat ja äänestävät asialistalla olevista asioista sekä muista esiin nousevista ajankohtaisista kysymyksistä. Parlamentit laativat täysistunnoista pöytäkirjat sekä julkaisevat niin pöytäkirjat kuin työskentelyn pohjalla olevat asiakirjat yleisölle saataviksi. Parlamenttien työn julkisuus ja läpinäkyvyys on tärkeää, jotta äänestäjät, media, tutkijat ja myös parlamentit itse voivat tarkastella

20 SPARQL. ks. `https://www.w3.org/TR/sparql11-query/`, on semanttisen webin standardi kyselykieli verkkomuotoista RDF-dataa varten, joka muistuttaa perinteisten tietokantojen SQL-kieltä.

21 CSV-muotoa käytetään taulukkomuotoisen datan esitysformaattina esimerkiksi taulukkolaskennassa.

22 `https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/helsinki-digital-humanities-hackathon-2021-dhh21`

23 `https://creativecommons.org/licenses/by/4.0/deed.fi`

päätöksenteon vaiheita sekä kansanedustajien esittämiä näkemyksiä ja toimintaa työssään lainsäätäjänä.

Parlamenttiaineistoja on saatettu viime vuosikymmeninä digitaaliseen muotoon; parlamenttiaineistojen on arvioitu olevan digitoitujen lehtiaineistojen ohella yleisimmin digitoituja digitaalisten ihmistieteiden data-aineistoja (Andrushchenko ym., 2021). Aineistojen digitointi liittyy kahteen laajempaan kehityskulkuun. Ensinnäkin parlamentit itse, kuten yleisemmin valtioiden hallinnot, ovat ottaneet käyttöön digitaalisia tiedonhallinnan järjestelmiä. Tämän myötä parlamenttien toimintaa koskevat dokumentit laaditaan ja julkaistaan sähköisesti parlamenttien omien aineistopalveluiden kautta. Toinen kehityskulku on, että vanhempia, painettuja parlamenttiaineistoja on digitoitu samaan tapaan kuin muutakin kansallista kulttuuriperintöaineistoa. Painettuja aineistoja ovat digitoineet parlamentit itse sekä myös kulttuuriperintöorganisaatiot ja erilaiset tutkimushankkeet. Esimerkiksi Ruotsin kuninkaallinen kirjasto on digitoinut painetut valtiopäiväasiakirjat aina vuodesta 1521 vuoteen 1970[24]. Kokoelmaa täydentävät valtiopäivien omat digitaaliset aineistot[25] sekä mm. Uumajan yliopiston Westac-tutkimushankkeessa[26] tehtävä valtiopäiväaineistojen rikastus ja tutkimustyökalujen kehitystyö.

Digitaalisen muodon myötä parlamenttiaineistoja ja niiden käytettävyyttä on parannettu niin yleisöä kuin tutkimuskäyttöäkin silmällä pitäen. Parlamentit ja tutkimushankkeet ovat luoneet verkkosivuja, joiden kautta käyttäjien on helppo selata ja ladata aineistoja. Näitä ovat esimerkiksi Kanadan parlamenttiaineistoja digitoineen Lipad-hankkeen (Beelen ym., 2017) sivusto[27] tai Italian edustajainhuoneen historiaa (1848–2018) monipuolisesti esittelevä portaali[28].

Täysistuntopöytäkirjoista on muodostettu myös parlamenttikorpuksia, jotka mahdollistavat puheiden sisällön sekä niiden kielen piirteiden tutkimuksen (esim. Lapponi ym., 2018; ks. myös CLARINin luettelo parlamenttikorpuksista[29]). CLARIN-infrastruktuurin piirissä on kehitetty myös keskustelupöytäkirjoille tarkoitettua TEI-pohjaista Parla-CLARIN-skeemaa[30], joka tarjoaa yhteisen tavan nimetä ja luokitella pöytäkirjakorpuksien sisältöä (esim. Pancur ym., 2020), ja jota hyödynnetään myös Semanttinen parlamentti -hankkeessa. Parla-CLARIN-yhteistyön myötä on syntynyt vertailevan tut-

24 https://riksdagstryck.kb.se/
25 http://data.riksdagen.se/
26 https://www.westac.se/
27 https://lipad.ca/
28 https://storia.camera.it/
29 https://www.clarin.eu/resource-families/parliamentary-corpora
30 https://github.com/clarin-eric/parla-clarin

kimuksen ParlaMint-hanke[31], jossa tuodaan yhteen Parla-CLARIN-muotoisia kansallisia korpuksia. Parlamenttiaineistoja on muunnettu myös linkitetyn datan muotoon, mikä on myös Semanttinen parlamentti -hankkeen yksi päämääristä. Esimerkkejä linkitetyn datan hankkeista ovat Euroopan Parlamentin aineistoja prosessoinut LinkedEP (van Aggelen, 2017), Italian parlamentti[32] tai Latvian parlamenttia koskeva LinkedSaeima-hanke (Bojārs ym., 2019).

Myös Suomessa valtiopäiväaineistoja on digitoitu eri yhteyksissä. Eduskunnan aineistot ovat kuitenkin hieman hankalasti käytettävissä, sillä niitä on tuotettu erikseen eri ajanjaksoista ja tallennettu eri muotoihin (Sinikallio ym. 2021). Eduskunnalla itsellään on avoimen datan aineistopalvelu, josta voi selata tai hakea kyselyillä uusimpia sähköisenä syntyneitä aineistoja. Lisäksi eduskunta on digitoinut ennen vuotta 2000 julkaistut painetut valtiopäiväpöytäkirjat ja -asiakirjat PDF-muotoisiksi. Näihin historiallisiin aineistoihin voi tehdä hakuja ja ne voi ladata niteittäin itselleen, mutta aineistojen käytettävyyttä heikentää niiden vaihteleva laatu ja kuvailutietojen puute (La Mela, 2020). Valtiopäiväkeskusteluista on julkaistu myös kielikorpuksia. Monipuolisin on FIN-CLARINin Kielipankin eduskuntakorpus, joka kattaa vuodet 2008–2016 ja sisältää kuvailutietoja ja esimerkiksi linkit alkuperäisiin täysistuntovideoihin, jotka on kohdistettu tekstiin automaattisin menetelmin (Eduskunta, 2017; Mansikkaniemi, Smit ja Kurimo, 2017). Tämä Eduskunnan täysistunnot -korpus on hyödynnettävissä Kielipankin Korp-palvelussa[33] (Lennes, 2019). Voices of Democracy -hanke on tuottanut tutkimuskorpuksen, joka sisältää kieliopillisesti ja puheenvuorojen tietojen osalta annotoidut täysistuntojen pöytäkirjat (1980–2018) sekä eduskunnan tekemiä veteraanikansanedustajien haastatteluja vuodesta 1988 (Andrushchenko ym., 2021). Eduskunnan keskustelupöytäkirjat vuosilta 1991–2015 löytyvät myös kansainvälisestä Harvard Parlspeech -korpuksesta (Rauh, 2017), mutta tässä korpuksessa olemme havainneet aukkoja kattavuuden osalta.

Digitoidut parlamenttiaineistot tarjoavat monenlaisia näkökulmia eri tutkimusaiheisiin, ja aineistoja on hyödynnetty monipuolisesti eri tutkimusaloilla esimerkiksi kielitieteessä, politiikan tutkimuksessa, mediatutkimuksessa, taloustieteessä ja historiantutkimuksessa. Keskeisin tutkimusaineisto ovat parlamenteissa käydyt keskustelut, joiden kautta voidaan tarkastella niin kieltä itseään kuin yhteiskunnallisia ilmiöitä laajasti. Metatietojen avulla on mahdollista jäsentää puhetta edelleen esimerkiksi puolueiden, sukupuolen tai

31 https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora

32 http://data.camera.it

33 http://korp.csc.fi

ammattiryhmien välillä. Luke Blaxill ja Kaspar Beelen (2016) ovat tarkastelleet naisten parlamentissa pitämien puheiden sisältöä sekä sukupuolen merkitystä edustajien puheessa Britannian parlamentin aineistojen avulla. Tutkimuksen kirjosta voidaan mainita parlamenttikeskustelujen teemojen tai niissä esiintyvien käsitteiden analyysi (esim. Quinn ym., 2010; Baker, Brezina & McEnery, 2017; Guldi, 2019; Ihalainen & Sahala, 2020; Kettunen & La Mela, 2021) ja puolueiden tai kansanedustajien kielen sekä näkemysten tutkimus (esim. Abercombie & Batista-Navarro, 2020; Magnusson ym., 2018). Parlamenttikeskusteluja on hyödynnetty myös muun muassa käännöstutkimuksessa esimerkiksi Euroopan parlamentin keskusteluista muodostetun EuroParl-korpuksen myötä[34].

Eduskunnan digitoituja aineistoja on hyödynnetty jonkin verran digitaalisten ihmistieteiden ja politiikan sekä yhteiskuntatieteiden tutkimuksessa. Matti La Mela (2020; myös Kettunen & La Mela, 2021) on tutkinut jokamiehenoikeuden käsitteen historiaa eduskunnan digitoitujen pöytäkirjojen avulla sekä tarkastellut PDF-muotoisten pöytäkirjojen laatua. Eduskunnan digitoituja pöytäkirjoja on hyödynnetty kieliteknologian menetelmien, tässä tapauksessa suomenkielisen semanttisen merkitsijän (*Finnish Semantic Tagger*) kehittämisessä (Kettunen & La Mela, 2021). Samoin Andrushschenko ym. (2021) ovat käyttäneet kieliopillisesti jäsennettyä korpustaan ja tätä varten kehitettyä hakutyökalua eri tutkimustapauksissa eduskuntakeskustelujen järjestämiseen ja analysointiin. Salla Simola (2020) on tarkastellut poliittisen puheen eroja puolueiden välillä koko eduskunnan ajalta 1907–2018, jota varten hän muodosti erillisen keskustelut ja puhujien tiedot yhdistävän tutkimusaineiston. Kimmo Makkonen ja Petri Loukasmäki (2019) ovat tutkineet eduskunnassa vuosina 1999–2014 pidettyjä täysistuntopuheita ja niiden sisältöjä monipuolisesti aihemallinnuksen (*topic modeling*) avulla. FIN-CLARINin Eduskunta-korpusta ovat hyödyntäneet esimerkiksi Ella Lillqvist ym. (2020) julkisesta velasta käytyä keskustelua koskevassa tutkimuksessaan. Näitä eri aineistoversioita käyttäviä tutkimuksia yhdistävät niin aineistojen laadun kuin korpusten vuosirajausten asettamat rajoitukset. Parlamenttisampo vastaa tähän ongelmaan ja parantaa eduskunta-aineiston käytettävyyttä niin tutkijoille kuin muillekin käyttäjäryhmille.

34 https://www.statmt.org/europarl/

Eduskunnan puheet 1907–2021 datana

Eduskunnan puheet -tietämysgraafi S-KG sisältää kaikista Suomen eduskunnan valtiopäivien täysistuntojen pöytäkirjoista kerätyt puheenvuorot aina vuodesta 1907 alkaen (Sinikallio ym., 2021). Aineiston keskeinen yksikkö on puhe (speech), joka kattaa keskustelupuheenvuorojen lisäksi puhemiesten istuntoa ohjaavat kommentit sekä kaikki muut istunnon osallistujien erikseen kirjatut kommentit esimerkiksi äänestyskäytäntöihin liittyen. Yksittäisiä puheita on aineistossa tällä hetkellä lähes miljoona kappaletta.

Tuotettu aineisto on kokonaisuudessaan olemassa kahtena eri versiona:

1. Parla-CLARIN-skeeman mukaisena XML-aineistona
2. Semanttinen parlamentti -hankkeessa kehitetyn puheskeeman mukaisena tietämysgraafina (S-KG)

Puheskeemassa käytetty Resource Description Framework (RDF)[35] on semanttisen webin perustan muodostava yksinkertainen graafimuotoinen tietomalli, jossa tieto esitetään verkon solmujen ja näiden välisten suunnattujen kaarien avulla. Verkkorakenteen semantiikka voidaan määritellä predikaattilogiikan avulla. W3C:n RDF-standardista ja siihen perustuvista muista tietomalleista ja kielistä on muodostunut keskeinen osa WWW:n nykyistä standardipinoa, jossa tiedon esittämisen kohteena on perinteisten syntaktisten XML-standardien yläpuolella olevat semanttiset rakenteet. Semanttisen webin teknologiaa on esitelty tarkemmin suomeksi teoksessa Hyvönen (2018).

RDF-muotoinen graafi S-KG on muodostettu täysistuntojen puheenvuoroista ja se linkittyy Kansanedustajien prosopografiseen tietämysgraafiin P-KG (Leskinen ym., 2021). Esimerkiksi puheiden pitäjät ja heidän edustamansa puolueet ovat P-KG-graafissa kuvattuja resursseja[36].

Aineisto on koottu useammasta eri datalähteestä ja alkuformaatista:

1. Valtiopäivistä 1907 valtiopäivien 1999 puoliväliin asti pöytäkirjat ovat saatavilla vain skannattuina, PDF-tiedostoon koottuina kuvina. Tämä aineisto muunnettiin ensin tekstintunnistusmenetelmillä (OCR) koneluettavaan muotoon. Muunnettuun aineistoon tehtiin pieniä manuaalisia korjauksia aineiston luotettavuuden parantamiseksi (Sinikallio ym., 2021).
2. Valtiopäivien 1999 puolivälistä valtiopäivien 2014 loppuun aineisto

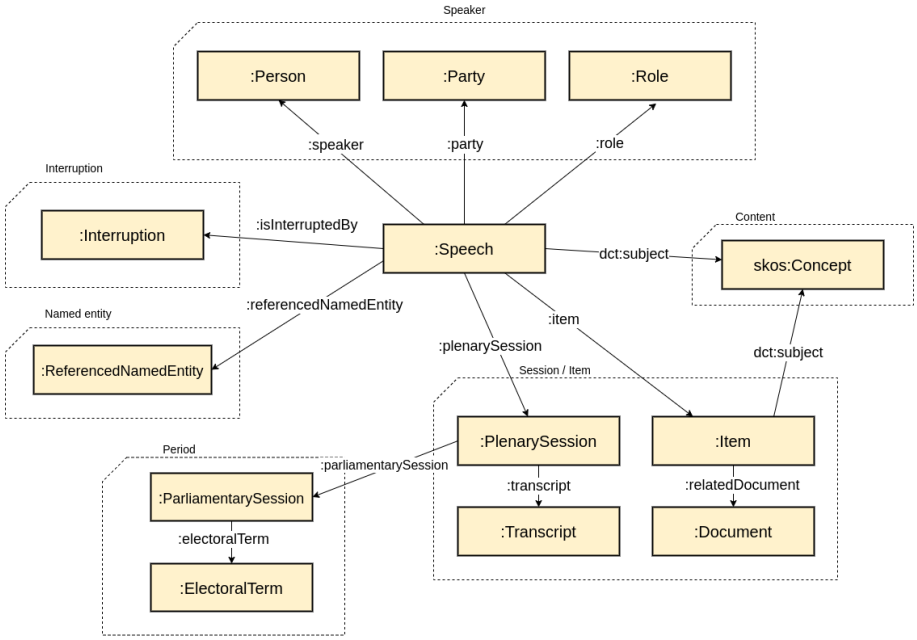
35 https://www.w3.org/RDF/

36 Resurssi (resource) on semanttisessa webissä objekti tai asia (esimerkiksi henkilö tai puhe), jolla on yksilöivä tunniste ja joka voidaan kuvata siihen liittyen ominaisuuksien ja toisten resurssien ja datan avulla.

kerättiin HTML-muotoisena eduskunnan verkkosivuilta[37].

3. Valtiopäivistä 2015 eteenpäin pöytäkirjat ovat saatavilla eduskunnan avoimen datan palvelun rajapinnasta[38] XML-muotoisena.

Kaikki aineistot muunnettiin alkuperäisformaatista ensin yhtenäiseen mm. taulukkolaskennassa käytettyyn CSV-muotoon, yksi puhe per taulukon rivi, ja tästä lopullisiin Parla-CLARIN- ja RDF-formaatteihin.



Kuva 2. Täysistuntojen puheenvuorojen tietämysgraafin keskeiset luokat ja näiden väliset suhteet. Nimiavaruus skos *viittaa Simple Knowledge Organization System (SKOS)[39] -tietomalliin ja* dct *Dublin Core Terms -metatietomäärittelyyn[40].*

Aineisto sisältää varsinaisten puheiden lisäksi kaiken oleellisen metadatan mitä pöytäkirjoihin on liitetty, kuten mahdolliset välihuudot, tietoja istunnosta, jossa puhe pidettiin (aika, päivämäärä, järjestysnumero jne.), puhujan tietoja (nimi, rooli, puolue) sekä mahdollinen keskustelun aihe (ts. asiakohta) sekä sitä alustavat dokumentit (esim. valiokunnan mietintö). Saatavilla olevan

37 https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx
38 https://avoindata.eduskunta.fi/#/fi/home
39 https://www.w3.org/2009/08/skos-reference/skos.html
40 https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

metadatan määrä vaihtelee suuresti alkuperäisestä formaatista riippuen. Muunnosprosessin aikana näitä tietoja on täydennetty tarpeen ja mahdollisuuksien mukaan monin eri keinoin, esimerkiksi hyödyntämällä kansanedustajien prosopografista tietämysgraafia P-KG puhujien tietojen osalta. RDF-muotoista S-KG tietämysgraafia on lisäksi rikastettu poimimalla ja linkittämällä puheista nimetyt entiteetit sekä asiasanoittamalla automaattisesti puheet ja tunnistamalla keskustelujen aiheita.

Kuvassa 2 näkyvät puhegraafin keskeiset luokat ja ominaisuudet. Puheeseen (:Speech) liittyy aina puhuja, tämän puolue ja rooli (luokat *:Person*, *:Party* ja *:Role*). Samoin puhe on aina osa jotain täysistuntoa (*:PlenarySession*), josta on olemassa lähteenä käytetty pöytäkirja (*:Transcript*). Täysistunto kytkeytyy aina tiettyihin valtiopäiviin (*:ParliamentarySession*), joka puolestaan kuuluu tiettyyn vaalikauteen (*:ElectoralTerm*). Puheesta voi olla tiedossa aihe, eli istunnon asiakohta (*:Item*), ja asiakohtaan liittyy useimmiten dokumentteja (*:Document*). Puheessa voi olla pöytäkirjassa siihen upotettuja keskeytyksiä, kuten välihuutoja. Nämä on poimittu myös omiksi kokonaisuuksikseen (*:Interruption*). Puheeseen voi liittyä siinä mainittuja nimettyjä entiteettejä (*:ReferencedNamedEntity*) sekä sen ja aiheen automaattisesti tuotetut asiasanat (*skos:Concept*).

Puhegraafia S-KG voidaan jatkossa päivittää helposti uusilla puheilla ajamalla muunnosprosessi uusilla aineistoilla.

Kansanedustajien verkostot 1907- datana

Datajulkaisu kansanedustajista (P-KG) kattaa kaikki Suomessa toimineet kansanedustajat tietämysgraafina (Leskinen ym., 2021). Sen ytimenä on RDF-tietomalliin tehty muunnos eduskunnan avoimen datan palvelusta[41] ladattavasta XML-muotoisesta datasta. Data sisältää perustietojen kuten synnyin- ja kuolinaikojen ja -paikkojen lisäksi yksityiskohtaisempaa tietoa henkilöiden elämäntapahtumista kuten opiskelusta, työelämästä, poliittisesta urasta tai heidän kirjallisista julkaisuistansa. Eduskunnan avoin data -tietolähdettä on täydennetty valtioneuvoston sivuilta[42] ja Wikidatasta louhituilla tiedoilla: tietämysgraafiin on lisätty kansanedustajien lisäksi 200 muuta poliittisen historian kannalta merkittävää henkilöä kuten presidenttejä, ministereitä ja oikeusasiamiehiä; esimerkiksi Mauno Koivisto on toiminut presidenttinä ja pääministerinä muttei koskaan kansanedustajana.

41 https://avoindata.eduskunta.fi/#/fi/dbsearch
42 https://valtioneuvosto.fi/hallitukset-ja-ministerit

Henkilöiden lisäksi graafi sisältää tietoa organisaatioista, ammateista ja viroista sekä paikoista. Organisaatioihin kuuluvat esimerkiksi puolueet, ministeriöt, eduskuntaryhmät, valiokunnat ja vaalipiirit sekä poliittisen toiminnan ulkopuolelta lähdeaineistosta tunnistetut koulut, järjestöt ja yritykset. Poliittiset organisaatiot on poimittu koneellisesti lähtöaineistosta ja ulkopuoleisten organisaatioiden tunnistamisessa on hyödynnetty KANTO-toimijaontologiaa[43]. Poliittiset virat, kuten *puhemies* tai *ulkoministeri* on generoitu lähtöaineistosta, muissa ammateissa ja oppiarvoissa on hyödynnetty AMMO-ammattiontologiaa[44] (Koho ym., 2018). Osa datasta on päätelty lähtötiedoista, esimerkiksi tiedot hallituspuolueista on luotu kulloisten ministerien puoluetietojen perusteella.

Tietomallissa käytetään Bio CRM[45] -pohjaista toimija–tapahtuma-mallia (Tuominen ym., 2018), joka on CIDOC CRM -standardin ja -ontologian[46] laajennus elämäkerrallisen tiedon esittämistä varten. Siinä henkilön elämä ja ura kuvataan häneen liittyvien yksittäisten tapahtumien sarjana, joihin hän osallistuu eri rooleissa, eri paikoissa ja eri aikoina; mallin mukaan henkilö liittyy roolinsa kautta tapahtumaan ja tapahtuma edelleen sisältää linkit vastaavaan ajanjaksoon sekä mahdollisiin paikkoihin ja organisaatioihin. Tämä ns. tapahtuma-perustainen (event-based) biografisen tiedon mallintamistapa on vastaava kuin esimerkiksi Sotasammossa[47] ja Biografiasammossa[48].

Linkitetyn datan palvelun käyttö

Edellä kuvattujen S-KG- ja P-KG-graafien julkaisualustana käytetty Linked Data Finland -palvelu tarjoaa linkitetyn datan julkaisemiseen ja käyttöön tarvittavat välineet ja palvelut, kuten IRI-tunnisteiden[49] uudelleenohjaukset, datan selausvälineen, datan latausmahdollisuuden (download), automaattisen dokumentoinnin välineitä ja ennen kaikkea SPARQL-palvelupisteen.

Keskeinen idea Parlamenttisammon datapalvelussa loppukäyttäjien kannalta on, että siinä käyttöön otettujen uusien tietomallien ja formaattien

43 https://www.kiwi.fi/display/Finto/KANTO+-+Kansalliset+toimijatiedot
44 http://light.onki.fi/ammo/fi/
45 https://seco.cs.aalto.fi/projects/biographies/
46 http://www.cidoc-crm.org/
47 https://seco.cs.aalto.fi/projects/sotasampo/
48 https://seco.cs.aalto.fi/projects/biografiasampo/
49 Semanttisen webin yksi keskeinen innovaatio on globaalisten yksilöivien verkkotunnisteiden käyttö, joiden avulla tietoalkioihin voidaan liittää metatietoa. IRI (Internationalized Resource Identifier) on yleinen tunnisteiden formaatti, jossa voidaan käyttää mm. ääkkösiä.

avulla data voidaan tarjota aiempaa rikkaammissa ja käyttökelpoisemmissa muodoissa ja tavoilla tutkijoita sekä sovellusten kehittäjiä varten. Parlamenttisampo-portaalin avulla tietoa voi hakea ja selata tutkijoiden ohella myös laajempi yleisö ilman ohjelmointitaitoa.

Aiemmat hakusovellukset eduskunnan puheaineistoille, kuten Kielipankin Korp (The Parliament of Finland, 2017) ja hakukoneet (Andrushchenko ym., 2021) perustuvat pohjimmiltaan tapauskohtaiseen perinteiseen tekstihakuun jäsennetystä kieliaineistosta. Haku kohdistuu vain pieneen osaan eduskunnan puheiden koko aikasarjasta. Tulosten tutkimiseen on tarjolla niukalti hakukoneeseen integroituja data-analyysiin liittyviä työvälineitä, kuten Kielipankin Korpin konkordanssianalyysi, jossa löydetyt sanat visualisoidaan niiden tekstikonteksteissa, ja hakutulosten statistiikat. Ajatuksena tällaisissa palveluissa on enempi tulosten tutkiminen lähiluvulla tai hakutulosten lataaminen johonkin ulkoiseen analyysijärjestelmään esimerkiksi CSV-muodossa. Parlamenttisammossa perinteisten käyttötapojen lisäksi tarjolla on standardinmukaisen SPARQL-kyselyrajapinnan tarjoamat uudet joustavat mahdollisuudet 1) datan avaamiseen kaikkien hyödynnettäväksi, 2) älykkääseen hakuun, 3) data-analyysiin erilaisilla työkaluilla ja itse ohjelmoimalla sekä 4) ulkoisten sovellusten kehittämiseen SPARQL-rajapinnan avulla. Esimerkkinä SPARQL-rajapinnan hyödyntämisestä sovelluksina on Parlamenttisampo-portaali ja sen valmiit haku-, selailu- ja analyysitoteutukset. Uutena ideana tässä portaalissa on, että hakua voidaan tehdä tekstihaun ohella ontologioihin perustavana fasettihakuna ja analysoida tuloksia tähän saumattomasti integroitujen visualisointi- ja data-analyysin työkalujen avulla Sampo-mallin mukaisesti. Fasettihaku (Tunkelang, 2009; Tzitzikas ym., 2017) perustuu S. R. Ranganathanin jo 1930-luvulla ideoimaan fasettiluokituksen teoriaan kirjastotieteessä. Semanttiseen, ts. koneen “ymmärtämään”, dataan perustuva lähestymistapa mahdollistaa myös tekoälyyn perustuvaa tietämyksen automaattista muodostamista (knowledge discovery).

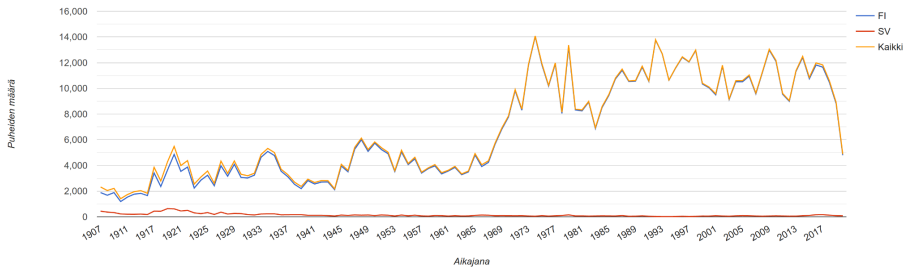
Seuraavassa esitellään tapoja hyödyntää Parlamenttisammon linkitetyn datan palvelua esimerkeillä havainnollistettuna sekä Parlamenttisampo-portaalin käyttöä. Varsinaisia parlamenttiaineistojen data-analyysien tuloksia tullaan julkaisemaan vasta myöhemmin niiden valmistuttua. Sampo-mallin lähestymistapaan liittyviä data-analyyttisiä mahdollisuuksia ja analyyseihin liittyviä haasteita humanistisessa tutkimuksessa on tarkasteltu julkaisussa (Tamper ym., 2021) liittyen Biografiasampoon, jossa puhetekstien sijaan tutkimuksen kohteena ovat Suomalaisen Kirjallisuuden Seuran Kansallisbiografian elämäkerrat.

Datan lataaminen perinteisiin työkaluihin

Yksinkertainen tapa tutkijalle on ladata aineistoja datapalvelusta paikallisesti käytettäväksi ja käyttää data-analyysiin perinteisiä analyyseissä käytettäviä ohjelmistoja ja työkaluja, esimerkiksi taulukkolaskimia CSV-muotoiselle datalle, R-ympäristöä[50] tilastollisiin analyyseihin tai Gephi-työkalua[51] verkostoanalyyseihin.

Parlamenttisammon tarjoama uudenlainen tapa on hyödyntää datapalvelun SPARQL-rajapintaa verkon kautta. Mahdollista on myös asentaa omalle koneelle uusi linkitetyn datan SPARQL-palvelinympäristö, esimerkiksi Fuseki[52], jota käytetään myös Linked Data Finland -palvelussa. Linked Data Finland -palvelun aineistot julkaistaan konttiteknologiaa hyödyntäen (Docker[53]), jolloin sekä datan että sen käsittelyssä tarvittavien eri ohjelmistojen muodostaman usein mutkikkaan kokonaisuuden asentaminen omalle koneelle on automaattista ja vaivatonta.

SPARQL-rajapinnan käyttäminen



Kuva 3. Puheiden määrä eri kielillä aikajanalla.

SPARQL-kieli on joustava tapa RDF-muotoisen datan kyselemiseen. Haun tulos esitetään taulukkomuodossa, jota voi tutkia sellaisenaan, visualisoida ja ohjelmoida sovelluskohtaisia analyysejä. Kuvassa 3 esitetään esimerkkinä visualisointi S-KG-graafin puheiden määrä kielen mukaan aikajanalla vuodesta 1907 vuoteen 2020 saakka. Kuvaajasta voidaan nähdä, kuinka puheiden määrä on muuttunut ajan saatossa. Suomenkielisiä puheita (kuvassa FI) on

50 https://www.r-project.org
51 https://gephi.org
52 https://jena.apache.org/documentation/fuseki2/
53 https://www.docker.com

pidetty selvästi eniten alusta asti. Alunperin ruotsinkielisiä puheita (kuvassa SV) on ollut enemmän kuin nykyisin, mutta niiden määrä jää hyvin vähäiseksi.

Kuvan grafiikka on muodostettu YASGUI-editorin[54] (Rietveld & Hoekstra, 2017) avulla, jolla voi näppärästi editoida SPARQL-kyselyitä, kohdistaa ne verkossa olevaan SPARQL-palvelupisteeseen ja visualisoida tuloksia työkaluun valmiiksi toteutettujen visualisointien avulla.

Kuvan 3 muodostamisessa käytetty SPARQL-kysely on esitetty alla:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX semparls: <http://ldf.fi/schema/semparl/>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX dct: <http://purl.org/dc/terms/>

SELECT ?year ( COUNT(?fin) as ?FI ) ( COUNT(?swe) as ?SV ) ( count(?-
documenturi) as ?Kaikki ) WHERE {
 ?documenturi a semparls:Speech .
 ?documenturi <http://purl.org/dc/terms/date> ?dateTime .
 BIND(STR(year(?dateTime)) as ?year)
 {
   BIND( <http://id.loc.gov/vocabulary/iso639-2/swe> as ?swe)
   ?documenturi dct:language ?swe .
 } UNION {
    BIND( <http://id.loc.gov/vocabulary/iso639-2/fin> as ?fin)
   ?documenturi dct:language ?fin .
 }
} GROUP BY ?year ORDER BY ASC(?year)
```

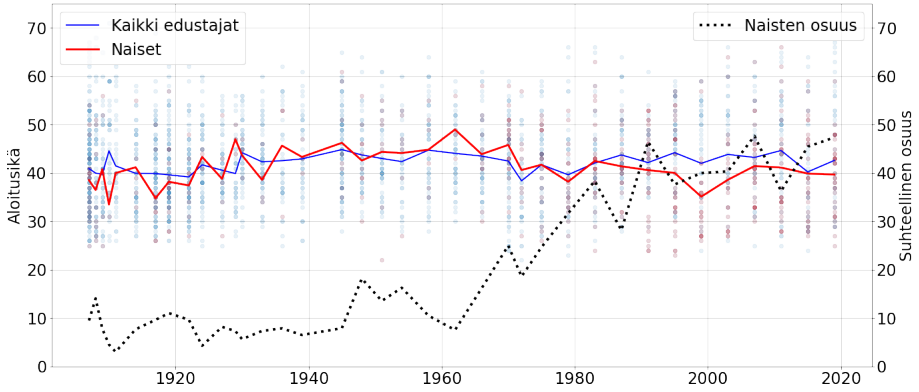
SPARQL-kielen tarkempi kuvaaminen ei tässä yhteydessä ei ole mahdollista, mutta pienellä harjoittelulla kielen voi oppia ilman varsinaista ohjelmointitaitoa. Kyseessä on varsin ilmaisuvoimainen ja joustava tapa hakea tietoa graafimuotoisesta datasta, ja se soveltuu digitaalisten ihmistieteiden tutkijoiden käytettäväksi (Hyvönen, 2018). Tässä SPARQL-kyselyssä esitellään aluksi käytettävät nimiavaruudet (PREFIX). Sen jälkeen tulevassa SELECT-kyselyssä haetaan kaikki puheet ja niiden kielet ?-alkuisten muuttujien avulla muodostetun graafihahmon avulla, jota sovitetaan palvelupisteen graafiin kaikilla mahdollisilla tavoilla. Vastaukseksi saadaan taulukko mahdollisista arvosijoituksista muuttujille. Tulokset luokitellaan lopuksi (GROUP

54 https://yasgui.triply.cc/

BY) kielen mukaan ryhmiksi, järjestetään vuoden mukaan (ORDER BY) ja lopuksi summataan (COUNT) kuinka paljon on puheita suomeksi, ruotsiksi ja yhteensä. Visualisoinnissa muuttuja ?year muodostaa X-akselin ja Y-akselilla ovat vuosittaiset numeeriset puheiden lukumäärät eri kielillä. Tässä kyselyssä on rajattu pois puheet, joilla ei ole kielikoodia. Puheiden kielen tunnistus on tehty koneellisesti ja toisinaan esimerkiksi OCR-virheet voivat vaikeuttaa kielen tunnistamista.

Aineistojen ohjelmallinen analyysi

Parlamenttisammon linkitettyä dataa voidaan tutkia laskennallisesti esimerkiksi Python-ohjelmointikielellä Google Colab[55] -ympäristössä. Käyttötapana on yksinkertaisen HTTP-protokollan käyttäminen ohjelmallisesti SPARQL-kyselyiden tekemiseen ja vastausten analysointi ja visualisointi käytetyn kielen tai ympäristön tarjoamilla välineillä, esimerkiksi Python-kirjastoilla.

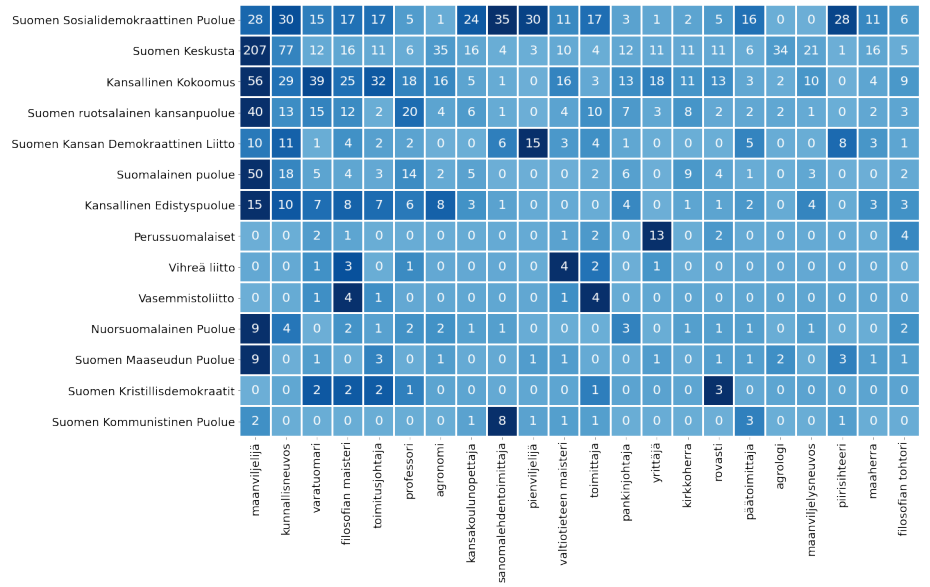


Kuva 4. Uusien kansanedustajien aloitusikä vuosittain ja naisten suhteellinen osuus.

Näin voidaan havainnollistaa esimerkiksi aineistojen ajallisia muutoksia tai ominaisuuksien välisiä korrelaatioita taulukoiden ja verkostojen avulla. Kuvassa 4 on esimerkkinä esitetty ensimmäisen kerran kansanedustajiksi valittujen henkilöiden aloitusikä vuosittain (Leskinen ym., 2021). Kuvassa sininen, yhtenäinen viiva esittää kaikkien kansanedustajien iän, naisten ikä on esitetty punaisella. Kuvasta voidaan nähdä, että aloitusikä on säilynyt lähes vakiona koko eduskunnan toiminnan ajan, mutta toisaalta vuoden 1980 jälkeen

55 https://colab.research.google.com/

naiset ovat olleet jonkin aikaa miehiä nuorempia aloittaessaan kansanedustajana. Naisten suhteellinen osuus on esitetty mustalla pisteviivalla. Ennen 1960-lukua osuus on pysynyt keskimäärin kymmenessä prosentissa, mutta kasvanut sittemmin 30–50 prosenttiin. Kuvan grafiikka on toteutettu Google Colab -dokumentilla data-analyysiin tarkoitettujen valmiiden Python-kirjastojen avulla.



Kuva 5. Taulukko ammatin ja puolueen välisistä korrelaatioista (Leskinen ym., 2021).

Kuvassa 5 esitetään vastaavalla tavalla muodostettu taulukkomuotoinen visualisointi puolueen ja edustajan ammatin välisestä korrelaatiosta. Puolueista ja ammateista on taulukkoon poimittu kaikkein yleisimmät eduskunnan koko toiminta-ajalta. Puolueet on esitetty taulukon vaakariveillä ja kunkin ammatin edustajien lukumäärä on merkitty ammattia vastaavan pystyrivin kohdalle. Näistä suosituin ammatti on merkitty tummimmalla taustavärillä. Lukumääräisesti eniten esiintyvä ammatti on maanviljelijä, jolla on eniten edustajia esimerkiksi keskustalla, kokoomuksella ja ruotsalaisella kansanpuolueella. Toisaalta esimerkiksi yrittäjä on ollut yleinen ammatti perussuomalaisilla tai rovasti kristillisdemokraateilla.

Parlamenttisampo-portaali

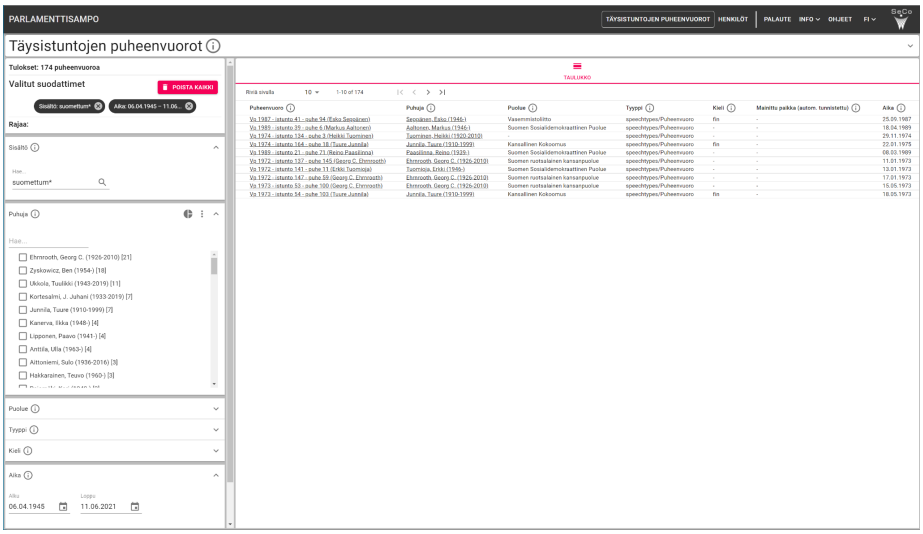
Edellä on kuvattu FinnParla-datapalvelun hyödyntämistä SPARQL-rajapinnan kautta (vrt. kuva 1). Myös hankkeessa kehitettävä semanttinen Sampo-portaali perustuu SPARQL-rajapinnan käyttöön Sampo-mallin periaatteiden mukaisesti Sampo-UI työkalulla toteutettuna. Portaali tarjoaa joukon erillisiä, mutta toisiinsa linkittyviä sovellusnäkymiä, joiden kautta palvelun sisältämää dataa voidaan hakea ja selata semanttista fasettihakua hyödyntäen temaattisesti. Esittelemme seuraavassa lyhyesti Parlamenttisampo-portaalin fasettihaun käyttöä esimerkin avulla.

Kuvassa 6 on esitetty Parlamenttisammon pääsivu (landing page), jossa on tässä vaiheessa kaksi temaattista sovellusnäkymää: puhedataan perustuva "Täysistuntojen puheenvuorot" ja politiikkojen verkostoihin perustuva "Henkilöt". Kuvaketta klikkaamalla avautuu vastaava sovellusnäkymä.

Kuva 6. Parlamenttisammon kehitteillä olevan portaalin pääsivu, jossa on kaksi temaattista, toisiinsa linkittyvää sovellusnäkymää. Uusien näkymien lisääminen on mahdollista dataa uudelleen käyttämällä.

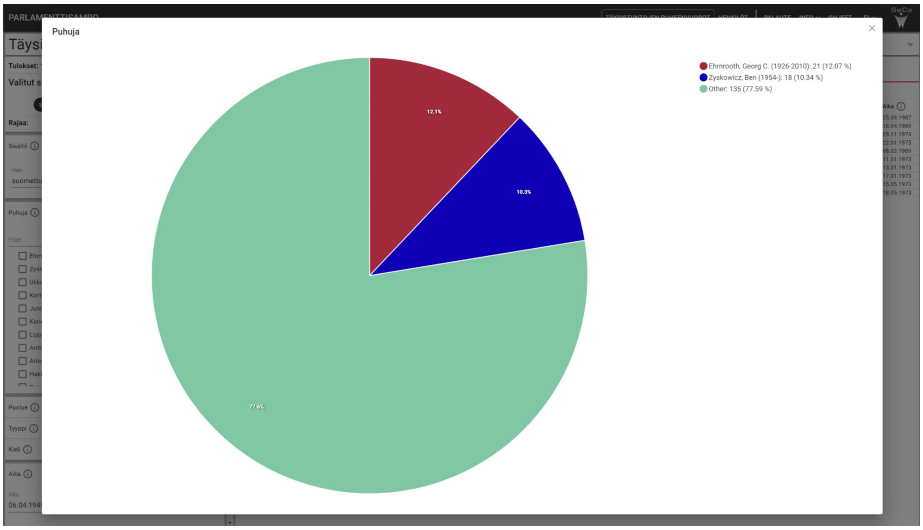
Kuvassa 7 käyttäjä on valinnut Täysistuntojen puheenvuorot -näkymän, jossa näkyy vasemmalla hakufasetit Sisältö, Puhuja, Puolue, (puheen) Tyyppi, Kieli ja Aika. Hakutulos, ts. löydetyt puheenvuorot, on esitetty taulukkomuodossa oikealla. Tuloksena on tässä tapauksessa 174 puhetta, jotka on sivutettu 10 tulosta per sivu. Käyttäjä on kirjoittanut Sisältö-tekstifasettiin kyselyn "suomettum*", jolloin hakutulokseen on suodattunut vain sellaiset puheet, joissa esiintyy sana "suomettuminen" eri muodoissaan, sillä jokerimerkki "*" sopii mihin tahansa merkkijonoon. Hän on myös rajannut tulosta

Aika-fasetilla vain puheisiin, jotka on pidetty 6.4.1945 alkaen, jolloin eduskunta alkoi kokoontua toisen maailmansodan jälkeen. Fasettihaussa rajaukset voidaan tehdä joustavasti missä järjestyksessä tahansa, ja hakukone laskee jokaiselle seuraavalla fasettivalinnalle osumaluvun (hit count), joka kertoo, kuinka monta tulosta tulosjoukkoon saataisiin, jos seuraavaksi tehdään kyseessä oleva valinta. Esimerkiksi klikkaamalla seuraavaksi Puhuja-fasetista "Junnila, Tuure (1910-199) [7]" löytyvät Tuure Junnilan seitsemän puhetta, joissa on mainittu suomettuminen. Valintafasetit on muodostettu automaattisesti sisällönkuvailussa käytetyn, FinnParlan taustalla olevan parlamenttiontologian ja datagraafien avulla.



Kuva 7. Suomettumiseen liittyvän puhejoukon rajaaminen fasettihaulla.

Osumaluvun avulla käyttäjää voidaan ohjata valintoihin, jotka eivät johda umpikujiin, jossa hakujoukko on tyhjä. Lisäksi osumaluvut tarjoavat mahdollisuuden tarkastella tulosjoukkoa tilastollisesti eri fasettien suhteen. Esimerkiksi Puhuja-fasetissa olevaa piirakkasymbolia klikkaamalla avautuu kuvan 8 mukainen piirakkakuvio, josta selviää kuinka monessa eri puheessaan eri puhujat mainitsevat suomettumisen. Voittajaksi tässä kisassa kirii Georg C. Ehrnrooth (21 puhetta) ja hopealle Ben Zyskowicz (18 puhetta). Muiden puhujien piirakkasiivut ovat pienempiä ja näkyvät siksi visualisoinnissa yksinkertaisuuden vuoksi koostettuna yhteen yhteiseen palaan (Other). Parametria muuttamalla pienemmätkin osuudet saataisiin näkyviin.



Kuva 8. Suomettuminen-sanan eri muodoissaan sisältävät puheet piirakkakaaviona Puhuja-fasetin kansanedustajien jakauman mukaan laskettuina.

Parlamenttisammon sovellusnäkymiin integroidaan jatkossa Sampo-mallin mukaisesti joukko valmiiksi toteutettuja data-analyysin välineitä ja visualisointeja, vastaavantyyppisiä kuin kuvissa 3–5 ja 8 on esitetty, joilla voidaan tutkia tarkemmin fasettihaulla rajattuja tulosjoukkoja. Työkalut ja visualisoinnit löytyvät jatkossa kuvan 7 taulukkovisualisoinnin rinnalta omilla välilehdillä samaan tapaan kuin esimerkiksi Akatemisammon käyttöliittymässä (Hyvönen ym., 2021b); molempien portaalien toteutuksessa uudelleenkäytetään Sampo-UI-kehysjärjestelmän (Ikkala ym., 2021) komponentteja. Näiden työkalujen ja visualisointien kautta projektissa tutkitaan tekoälyn mahdollisuuksia tietämyksen muodostamisessa digitaalisten ihmistieteiden tutkimuksessa: miten Parlamenttisampo voisi esimerkiksi avustaa tutkijaa tutkimusongelmien automaattisessa ratkonnassa, ratkaisujen selittämisessä tai uusien tutkimusongelmien haussa? (Hyvönen, 2020)

Uusia mahdollisuuksia eduskuntatutkimuksessa

Politiikan tutkimuksen piirissä parlamenttipuhetta pidetään merkittävänä poliittisen viestinnän ja poliittisen kamppailun muotona. Parlamenttipuhe ei ole mitä tahansa puhetta, vaan sillä on oma rakenteensa ja omat sääntönsä, jotka samalla heijastelevat parlamentin yleistä asemaa. Tämän lisäksi parlamenttipuhe on poliittisen kamppailun väline, jonka avulla voidaan tuoda

näkyviksi kilpailevia tavoitteita, haastaa vastapuolen näkemyksiä sekä avata lukkiutuneita asetelmia. Puhe parlamentissa on siten aina myös poliittinen teko, jossa käytetyt sanat ovat politiikanteon aseita, ja jotka kertovat paitsi käsittelyssä olevista asioista, myös paljastavat puhujien erilaisia positioita, arvoasetelmia sekä näkökantoja. (Palonen, 2005)

Perinteisesti parlamenttipuhetta on meillä ja maailmalla tutkittu lähiluennalla ja hyödyntämällä sisällönanalyysiä, diskurssianalyysiä tai erilaisia retoriikan tutkimuksen menetelmiä. Digitalisaatio on kuitenkin työntynyt myös tälle perinteiselle tutkimusalueelle yhä voimakkaammin, kun eri maissa parlamenttikeskusteluihin liittyvää dataa on alettu tarjota avoimen datan muodossa yhä suurempina määrinä. Suomen eduskunnan osalta valtiopäiväasiakirjojen digitointi on edennyt kohtuullisella nopeudella, ja osa aineistoista on ollut tarjolla myös eduskunnan avoimen datan palvelun kautta. Aineistojen saatavuus ja myös tarjolla olevien aineistojen laatu on parantunut viime vuosina, mutta edelleen erot esimerkiksi Ruotsin, Ison-Britannian tai Saksan vastaaviin aineistoihin ovat huomattavat.

Semanttinen parlamentti -hanke on merkittävä askel eduskunnassa käytyjen täysistuntokeskustelujen hyödyntämisessä osana ihmistieteiden tutkimuskenttää. Vaikka aineistot ovat toki koko ajan olleet tutkijoiden käytettävissä manuaalisesti sekä joidenkin vuosien ajan myös sähköisesti PDF-muotoon digitoituna, nyt valmisteltava koneluettava aineistokorpus yhdessä myöhemmin lanseerattavan Parlamenttisampo-portaalin kanssa kytkevät eduskunnan täysistuntokeskustelut sekä muut avoimet aineistot osaksi digitaalisten ihmistieteiden kenttää ja kansallista tietoinfrastruktuuria. Käytännössä tämä tarkoittaa esimerkiksi politiikan tutkijoille, historioitsijoille ja kielitieteilijöille mahdollisuuksia louhia, mallintaa, analysoida ja visualisoida parlamenttipuhetta eksploratiivisen tutkimuksen keinoin hyödyntämällä valtavaa aineistokorpusta, joka kattaa koko modernin eduskuntamme toiminta-ajan vuodesta 1907 lähtien.

Mahdollisuus eksplorativiiseen data-analyysiin avaa aivan uusia mahdollisuuksia ja näkökulmia parlamenttipuheen tutkimukselle. Perinteisessä lähiluennassa (close reading) tutkija on pakotettu rajaamaan aineistoa jo sen keruuvaiheessa hyvinkin voimakkaasti, mikä useimmiten tapahtuu joko ajallisen tai temaattisen rajauksen kautta – siis joko fokusoitumalla rajattuun ajanjaksoon tai rajattuihin teemoihin. Digitaaliset menetelmät mahdollistavat aineiston tarkastelun ilman tällaisia esirajauksia, mikä mahdollistaa aineiston tutkimisen hyödyntämällä esimerkiksi täys- tai puoliautomaattisen luokittelun menetelmiä. Näin voi olla mahdollista löytää esimerkiksi uusia teemoja ja aihepiirejä, jotka ovat jääneet tutkimuksissa aiemmin sivuun (esim. Mimno, 2012; Tangherlini & Leonard, 2013). Toisaalta ilman voimakkaita esioletuksia

tapahtuva aineiston etäluenta ja luokittelu mahdollistaa myös aiempien tutkimustulosten kriittisen tarkastelun, kun etäluennan tuottamia aiheita tai teemoitteluja voidaan verrata muilla menetelmillä saatuihin tuloksiin (esim. Ylä-Anttila ym., 2018).

Toinen esimerkki data-aineiston tarjoamista mahdollisuuksista on politiikan kieleen ja sen muutokseen pitkällä aikavälillä kohdistuva tutkimus (mm. DiMaggio ym., 2013; Jacobi ym., 2016; Purhonen & Toikka, 2016; Laaksonen & Nelimarkka, 2018; Törnberg & Törnberg, 2016, Mountford, 2018). Valtava aineistomassa tarjoaa mahdollisuuden kieliteknologisten menetelmien laajamittaiselle ja systemaattiselle soveltamiselle. Vaikka parlamenttipuhe on myös kielellisesti oma erityinen puhunnan muotonsa, myös parlamenttipuhe elää ajassa ja heijastelee siten sekä laajempaa kielellistä kehitystä että yhteiskunnallista keskusteluilmapiiriä ja siinä esiintyviä sanavalintoja (Makkonen & Loukasmäki, 2019). Samalla laaja aineisto tarjoaa mahdollisuuden tutkia kielenkäytön muutosta, siis esimerkiksi sitä, onko yhteiskunnallinen keskusteluilmapiiri polarisoitunut tai ”raaistunut”, kuten viime vuosina sekä poliitikot ja mediatoimijat ovat toistuvasti esittäneet.

Kolmas eduskunnan datan tarjoama mahdollisuus liittyy kielen käytön kytkemiseen laajemmin kielen käyttäjien muuhun sosiaaliseen kontekstiin kuten koulutukseen, ikään ja sosiaalisiin verkostoihin. Kieltä voidaan politiikan tutkimuksessa lähestyä myös sen oletuksen pohjalta, että kieli heijastelee aina myös käyttäjänsä laajempaa arvo- ja aatemaailmaa sekä hänen sosiaalista asemaansa ja kontekstiaan. Diskursiiviset koalitiot, joita voidaan rakentaa puhujien kielenkäyttöön perustuen, tarjoavat siten kiinnostavan mahdollisuuden irrottautua esimerkiksi puoluetaustan asettamasta viitekehyksestä ja suunnata analyyttistä katsetta kielen käytön kautta rakentuviin verkostoihin. Aiemmissa tutkimuksissa tämän tyyppisellä lähestymistavalla on pystytty liittämään asiantuntijoita erilaisiin ideologisiin positioihin heidän tekstiensä sisältöjä analysoimalla (Jelveh ym., 2014), mitä perusajatusta voi mielestämme hyvin soveltaa myös kansanedustajien luokitteluun.

Edellä on nostettu esille muutamia esimerkkejä, joiden kohdalla eduskuntadatan hyödyntäminen näyttäisi mahdollistavan eduskuntatutkimuksessa uusia merkittäviä tutkimuksellisia avauksia. Eksploratiivisen data-analyysin hengessä on kuitenkin syytä korostaa niitä, vielä tuntemattomia mahdollisuuksia, jotka vähitellen avautuvat, kun aineistoja tutkimalla ja analysoimalla tutkijat alkavat hahmotella uusia hypoteeseja ja tutkimuskysymyksiä. Suurten data-aineistojen potentiaali on niissä piilevässä mahdollisuudessa yllättyä, mikä toisaalta vaatii ennakkoluulotonta asennetta aineistoa kohtaan, toisaalta alleviivaa data-analyysin parissa työskentelevien tutkijoiden kasvavaa vastuuta. Kun tutkijan ei ole mahdollista enää tuntea käyttämäänsä aineistoa

perinpohjin, hänen tulee tuntea aineiston kohteena olevat ilmiöt perinpohjin. Vain siten on mahdollista arvioida, mitkä louhinnan, analyysin, mallinnuksen tai visualisoinnin kautta välittyvät havainnot ovat oikeasti merkityksellisiä. (Elo, 2016)

Kiitokset

Semanttinen parlamentti -hanke on osa Suomen Akatemian rahoittamaa DIGIHUM-ohjelmaa 2020–2022, joka toteutetaan Helsingin yliopiston digitaalisten ihmistieteiden keskuksessa HELDIG (koordinaattori), Aalto-yliopistossa ja Turun yliopiston Eduskuntatutkimuksen keskuksessa. Hanke liittyy myös esineellisen ja aineettoman kulttuurin yhdistämistä tutkivaan EU-projektiin InTaVia[56] sekä EU COST -hankkeeseen Nexus Linguarum[57], jossa tutkitaan linkitettyyn dataan perustuvaa lingvististä data-analyysiä. Hankkeessa käytetään Tieteen tietotekniikan keskuksen (CSC) laskentaresursseja.

Kirjallisuusviitteet

Abercrombie, G., & Batista-Navarro, R. (2020). Sentiment and Position-Taking Analysis of Parliamentary Debates: a Systematic Literature Review. *Journal of Computational Social Science, 3*, 245–70. https://doi.org/10.1007/s42001-019-00060-w

Abu-Salih, B. (2021). Domain-specific Knowledge Graphs: A survey. *Journal of Network and Computer Applications, 185*(1), July 2021. https://doi.org/10.1016/j.jnca.2021.103076

Van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2017). The Debates of the European Parliament as Linked Open Data. *Semantic Web, 8*(2), 271–281.

Andrushchenko, M., Sandberg, K., Turunen, R., Marjanen, J., Hatavara, M., Kurunmäki, J., . . . Nummenmaa, J. (2021). Using parsed and annotated corpora to analyze parliamentarians' talk in Finland. *Journal of the Association for Information Science and Technology*, 1–15. https://doi.org/10.1002/asi.24500

Baker, H., Brezina V., & McEnery T. (2017). Ireland in British parliamentary debates: plotting changes in discourse in a large volume of time-series corpus data. Teoksessa T. Säily, A. Nurmi, M. Palander-Collin & A. Auer (toim.), *Exploring future paths for historical sociolinguistics* (s. 83–107). John Benjamins.

Beelen, K., Thijm, T. A., Cochrane, C., Halvemaan, K., Hirst, G., Kimmins, M., . . . Whyte, T. (2017). Digitization of the Canadian Parliamentary Debates. *Canadian Journal of Political Science, 50*(3), 849–864. http://doi.org/10.1017/S0008423916001165

56 https://intavia.eu/
57 https://nexuslinguarum.eu/

Benoît, C., & Rozenberg, O. (toim.) (2020). *Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures*. Edward Elgar Publishing. https://doi.org/10.4337/9781789906516

Blaxill, L., & Beelen, K. (2016). A Feminized Language of Democracy? The Representation of Women at Westminster since 1945. *Twentieth Century British History, 27*(3), 412–449. https://doi.org/10.1093/tcbh/hww028

Bojārs, U., Darģis, R., Lavrinovičs, U., & Paikens, P. (2019). LinkedSaeima: A Linked Open Dataset of Latvia's Parliamentary Debates. Teoksessa M. Acosta, P. Cudré-Mauroux, M. Maleshkova, T. Pellegrini, H. Sack, Y. Sure-Vetter (toim.), *Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019* (s. 50–56). Lecture Notes in Computer Science, vol 11702. Springer. https://doi.org/10.1007/978-3-030-33220-4_4

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting Affinities between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding. *Poetics, 41*(6), 570–606. https://doi.org/10.1016/j.poetic.2013.08.004

Eduskunta (2017). Eduskunnan täysistunnot, Kielipankin Korp-versio 1.5 [tekstikorpus]. Kielipankki. http://urn.fi/urn:nbn:fi:lb-2019101621

Elo, K. (2016). Digitaalisen historiantutkimuksen kenttää louhimassa. Teoksessa K. Elo (toim.), *Digitaalinen humanismi ja historiatieteet* (Historia Mirabilis 12) (s. 11–35). Turun historiallinen yhdistys.

Gardiner, E., & Musto, R. G. (2015). *The Digital Humanities: A Primer for Students and Scholars*. Cambridge University Press. https://doi.org/10.1017/CBO9781139003865

Guldi, J. (2019). Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change. *Technology and Culture, 60*(1), 1–33. http://dx.doi.org.ezproxy.its.uu.se/10.1353/tech.2019.0000

Heath, T., & Bizer, C. (2011). *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool. https://doi.org/10.2200/S00334ED1V01Y201102WBE001

Hyvönen, E. (2018). *Semanttinen web. Linkitetyn avoimen datan käsikirja*. Gaudeamus.

Hyvönen, E. (2020). Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery. *Semantic Web – Interoperability, Usability, Applicability, 11*(1), 187–193. https://doi.org/10.3233/SW-190386

Hyvönen, E. (2021). Digital Humanities on the Semantic Web: Sampo Model and Portal Series. Vertaisarvioinnissa. https://seco.cs.aalto.fi/publications/2021/hyvonen-sampo-model-2021.pdf

Hyvönen, E., Tuominen, J., Alonen, M., & Mäkelä, E. (2014). Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. Teoksessa V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, & A. Tordai (toim.), *The Semantic Web: ESWC 2014 Satellite Events. ESWC 2014* (s. 226–230). Springer. https://doi.org/10.1007/978-3-319-11955-7_24

Hyvönen, E., Tamper, M., Ikkala, E., Koho, M., Leal, R., Kesäniemi, J., . . . Hietanen, A. (2021a). LawSampo Portal and Data Service for Publishing and Using Legislation and Case Law as Linked Open Data on the Semantic Web. April. https://seco.cs.aalto.fi/publications/2021/hyvonen-et-al-lawsampo-2021.pdf

Hyvönen, E., Leskinen, P:, Rantala, H., Ikkala, E., & Tuominen, J. (2021b). Akatemiasampo-portaali ja -datapalvelu henkilöiden ja henkilöryhmien historialliseen tutkimukseen. *Informaatiotutkimus, 40*(2), 28–56, https://doi.org/10.23978/inf.102656

Ihalainen, P., & Sahala, A. (2020). Evolving conceptualisations of internationalism in the UK Parliament: Collocation analyses from the league to Brexit. Teoksessa M. Fridlund, M. Oiva, & P. Paju (toim.), *Digital histories: Emergent approaches within the new digital history* (s. 199–219). Helsinki University Press. https://doi.org/10.33134/HUP-5-12

Ikkala, E., Hyvönen, E., Rantala, H., & Koho, M. (2021). Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web – Interoperability, Usability, Applicability*. Accepted. http://www.semantic-web-journal.net/content/sampo-ui-full-stack-javascript-framework-developing-semantic-portal-user-interfaces-0

Jacobi, C., van Atteveldt, W., & Welbers, K. (2016). Quantitative Analysis of Large Amounts of Journalistic Texts Using Topic Modelling. *Digital Journalism, 4*(1), 89–106. https://doi.org/10.1080/21670811.2015.1093271

Jelveh, Z., Kogut, B., & Naidu, S. (2014). Detecting Latent Ideology in Expert Text: Evidence from Academic Papers in Economics. Teoksessa *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (s. 1804–1809). http://emnlp2014.org/papers/pdf/EMNLP2014191.pdf

Kettunen K., & La Mela M. (2021, tulossa) Semantic tagging and the Nordic tradition of Everyman's rights. *Digital Scholarship in the Humanities*. Preprint-versio (huhtikuu 2021). https://seco.cs.aalto.fi/publications/2021/kettunen-lamela-dsh-2021.pdf

Koho, M., Gasbarra, L., Tuominen, J., Rantala, H., Jokipii, I., & Hyvönen, E. (2019). AMMO Ontology of Finnish Historical Occupations. Teoksessa *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH 19)* (s. 91–96). CEUR Workshop Proceedings, vol. 2375. http://ceur-ws.org/Vol-2375/

Laaksonen, S.-M., & Nelimarkka, M. (2018). Omat ja muiden aiheet: Laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta. *Politiikka, 60*(2), 132–147.

La Mela, M. (2020). Tracing the emergence of Nordic allemansrätten through digitised parliamentary sources. Teoksessa M. Fridlund, M. Oiva, & P. Paju (toim.), *Digital histories: Emergent approaches within the new digital history* (s. 181–197). Helsinki University Press. https://doi.org/10.33134/HUP-5-11

Lapponi, E., Søyland, M. G., Velldal, E., & Oepen, S. (2018). The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. *Lang Resources & Evaluation, 52*, 873–893. https://doi.org/10.1007/s10579-018-9411-5

Lennes, M. (2019). FIN-CLARIN and Language Bank Parliamentary Data. *Workshop "Digital Parliamentary Data and Research"*, Aalto-yliopisto, 3.5.2019. https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/workshop-digital-parliamentary-data-and-research

Leskinen, P., Hyvönen, E., & Tuominen, J. (2021). Members of Parliament in Finland Knowledge Graph and its Linked Open Data Service. April. *Proceedings of SEMANTiCS - In the Era of Knowledge Graphs, Amsterdam, Sept 6-9, 2021*, accepted. https://seco.cs.aalto.fi/publications/2021/leskinen-et-al-mps-2021.pdf

Lillqvist, E., Kavonius, I. K., & Pantzar, M. (2020). “Velkakello tikittää”: Julkisyhteisöjen velka suomalaisessa mielikuvastossa ja tilastoissa 2000–2020. *Kansantaloudellinen Aikakauskirja, 116*(4), 581–607.

Magnusson, M., Öhrvall, R., Barrling, K., & Mimno, D. (2018, April 4). Voices from the far right: a text analysis of Swedish parliamentary debates. https://doi.org/10.31235/osf.io/jdsqc

Makkonen, K., & Loukasmäki, P. (2019). Eduskunnan täysistunnon puheenaiheet 1999–2014: Miten käsitellä LDA-aihemalleja?. *Politiikka, 61*(2), 127–159. https://journal.fi/politiikka/article/view/77163

Mansikkaniemi, A., Smit, P., & Kurimo, M. (2017). Automatic Construction of the Finnish Parliament Speech Corpus. Teoksessa *Proc. Interspeech 2017* (s. 3762-3766). https://doi.org/10.21437/Interspeech.2017-1115

Martínez-Rodríguez, J.-L., Hogan, A., & López-Arévalo. I. (2020). Information extraction meets the Semantic Web: A survey. *Semantic Web – Interoperability, Usability, Applicability, (11)*2, 255–335, https://doi.org/10.3233/SW-180333

Mimno, D. (2012). *Topic Regression*. University of Massachusetts Amherst. https://scholarworks.umass.edu/open_access_dissertations/520

McCarty, W. (2005). *Humanities Computing*. Palgrave.

Mountford, J. B. (2018). Topic Modeling the Red Pill. *Social Sciences, 7*(3). https://doi.org/10.3390/socsci7030042

Noy, N., Gao, Y., Jain, A., Narayanan, A., Patterson, A., & Taylor, J. (2019). Industry-scale knowledge graphs: lessons and challenges. *Communications of the ACM*, July. https://doi.org/10.1145/3331166

Oksanen, A., Tuominen, J., Mäkelä, E., Tamper, M., Hietanen, A., & Hyvönen, E. (2019). Semantic Finlex: Transforming, Publishing, and Using Finnish Legislation and Case Law As Linked Open Data on the Web. Knowledge of the Law in the Big Data Age. Teoksessa G. Peruginelli, & S. Faro (toim.), *Frontiers in Artificial Intelligence and Applications*, vol. 317 (s. 212–228). IOS Press.

Palonen, K. (2005). Eduskunnasta puhekunnaksi? Parlamentarismi retorisena politiikkana. *Politiikka, 47*(2), 142–148.

Pancur, A., & Erjavec, T. (2020). The siParl corpus of Slovene parliamentary proceedings. Teoksessa *Proceedings of the Second ParlaCLARIN Workshop. Marseille, France, May 2020* (s. 28–34). European Language Resources Association. https://www.aclweb.org/anthology/2020.509parlaclarin-1.6

Purhonen, S. & Toikka, A. (2016). “Big Datan” haaste ja uudet laskennaliset tekstiaineistojen analyysimenetelmät: esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935-2015. *Sosiologia, 53*(1), 6–27.

Quinn, K. M., Monroe, B. L., Colaresi, M., Crespin, M. H., & Radev, D. R. (2010). How to Analyze Political Attention with Minimal Assumptions and Costs. *American Journal of Political Science, 54*, 209–228. https://doi.org/10.1111/j.1540-5907.2009.00427.x

Rauh, C., De Wilde, P., & Schwalbach, J. (2017). The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states (V1). Harvard Dataverse. https://doi.org/10.7910/DVN/E4RSP9

Rietveld, L., & Hoekstra, R. (2017). The YASGUI family of SPARQL clients. *Semantic Web – Interoperability, Usability, Applicability, 8*(3), 373–383. https://doi.org/10.3233/SW-150197

Simola, S. (2020). A Century of Partisanship in Finnish Political Speech. Julkaisematon käsikirjoitus. Osa väitöskirjaa *Essays in Labor and Political Economics*, Aalto-yliopisto. https://sites.google.com/site/sallasimolaecon/home/research

Sinikallio, L., Drobac, S., Tamper, M., Leal, R., Koho, M., Tuominen, J. . . . Hyvönen, E. (2021). Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup. *Proceedings, Language, Data and Knowledge (LDK 2021), Zaragoza, Spain, June, 2021*, accepted. https://seco.cs.aalto.fi/publications/2021/sinikallio-et-al-speeches-2021.pdf

Staab, S. & Studer, R. (toim.) (2009). *Handbook of Ontologies*. Springer. https://doi.org/10.1007/978-3-540-92673-3

Tamper, S., Leskinen, P., Hyvönen, E., Valjus, R., & Keravuori, K. (2021). Analyzing Biography Collections Historiographically as Linked Data: Case National Biography of Finland. *Semantic Web – Interoperability, Usability, Applicability*, accepted. http://semantic-web-journal.org/content/analyzing-biography-collections-historiographically-linked-data-case-national-biography

Tangherlini, T. R., & Leonard, P. (2013). Trawling in the Sea of the Great Unread: Sub-Corpus Topic Modeling and Humanities Research. *Poetics, 41*(6), 725–749. https://doi.org/10.1016/j.poetic.2013.08.002

Tunkelang, D. (2009). *Faceted Search. Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan-Claypool. https://doi.org/10.2200/S00190ED1V01Y200904ICR005

Tuominen, J., Hyvönen, E., & Leskinen, P. (2018). Bio CRM: A Data Model for Representing Biographical Data for Prosopographical Research. Teoksessa *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)* (s. 59–66). CEUR Workshop Proceedings, vol. 2119. http://ceur-ws.org/Vol-2119/paper10.pdf

Tzitzikas, Y., Manolis, N., & Papadakos, P. (2017). Faceted Exploration of RDF/S Datasets: a Survey. *Journal of Intelligent Information Systems, 48*(2), 329–364. https://doi.org/10.1007/s10844-016-0413-8

Törnberg, A., & Törnberg, P. (2016). Muslims in Social Media Discourse: Combining Topic Modeling and Critical Discourse Analysis. *Discourse, Context and Media, 13*, 132–142. https://doi.org/10.1016/j.dcm.2016.04.003

Verboven, K., Carlier, M., & Dumolyn, J. (2007). A Short Manual to the Art of Prosopography. Teoksessa *Prosopography Approaches and Applications. A Handbook* (s. 35–70). Unit for Prosopographical Research (Linacre College).

Ylä-Anttila, T., & Eranti, V. (2018). Aihemallinnuksesta kehysmallinnukseen. *Politiikka, 60*(2), 148–156.