

Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language

Eero Hyvönen^{1,2}, Petri Leskinen^{1,2}, Laura Sinikallio^{2,1}, Matti La Mela²,
Jouni Tuominen^{1,2}, Kimmo Elo³, Senka Drobac^{2,1}, Mikko Koho^{1,2}, Esko Ikkala¹,
Minna Tamper^{1,2}, Rafael Leal^{1,2} and Joonas Kesäniemi¹

¹*Semantic Computing Research Group (SeCo), Department of Computer Science, Aalto University, Finland*

²*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*

³*Centre for Parliamentary Studies, University of Turku, Finland*

Abstract

This paper introduces the system *ParliamentSampo – Parliament of Finland on the Semantic Web*, a Linked Open Data (LOD) service, data infrastructure, and semantic portal for studying Finnish political culture, language, and networks of the Members of Parliament (MP). The article presents the vision behind the system, the LOD service, and explores the possibilities to utilize it in research and application development. A knowledge graph of linked data has been created based on ca. 962 000 speeches in all plenary sessions of the Parliament of Finland in 1907–2021; the data is also available in XML format, utilizing the new international Parla-CLARIN format. For the first time, the entire time series of the Finnish parliamentary speeches has been converted into data and a data service in a unified format. In addition, the speeches have been interlinked with another knowledge graph created from the database of the MPs and enriched from other data sources into a broader ontology-based data service. The paper shows how the LOD service SPARQL endpoint can be used to research parliamentary culture, the use of political language, and networks of politicians through data analysis. The service endpoint can also be used to develop applications for different user groups without programming skills, such as the PARLIAMENTSAMPO semantic portal introduced in the paper, too. This application aims to make political decision making more transparent to the general public, media, politicians, and other end users.

Keywords

parliamentary studies, semantic portals, linked data, digital humanities

1. Introduction

The main tasks of parliaments are to enact new laws, oversee the work of the government, and decide on the state budget; how the parliament works in Finland is documented in [1]. Parliamentary data are used in many areas of research [2], as it provides a wealth of information on the state and functioning of democratic systems, political life and, more generally, language and culture. The most prominent part of the work of parliaments is the public plenary sessions, in which the Members of Parliament (MP) discuss and vote on issues on the agenda and other topics


Digital Parliamentary Data in Action (DiPaDa 2022) workshop, Uppsala, Sweden, March 15, 2022.

✉ eero.hyvonen@aalto.fi (E. Hyvönen)

🌐 <https://seco.cs.aalto.fi/u/eahyvone/> (E. Hyvönen)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

that arise. Parliaments draw up minutes of plenary sessions and make both the minutes and the documents on which they are based available to the public. Openness and transparency in the work of parliaments is important for the voters, media, researchers, and also parliaments themselves: based on open data, they can look at the decision-making stages, views, and actions expressed by parliamentarians in their work as legislators.

This paper argues, inspired by [3, 4], for publishing and using parliamentary data in Digital Humanities (DH) research based on Semantic Web (SW) technologies¹ and Linked Data (LD) [5]. The LD approach for Cultural Heritage [6] has arguably many advantages: 1) Linked data and ontologies [7] provide a framework for harmonizing heterogeneous distributed datasets and combining them into larger and richer entities. 2) The SW is based on the Predicate Logic [8], which provides an opportunity to enrich data by reasoning new information. 3) When the machine “understands” the content of the data, intelligent web services and data analyses can be implemented more easily. 4) Ready-made tools by other actors can be re-used for publishing, processing and analysing the standardized data; the wheel doesn’t need to be reinvented.

In this paper, we test and demonstrate the above arguments in DH research on parliamentary culture and language [2] by presenting the PARLIAMENTSAMPO system, a Linked Open Data (LOD) corpus and data service of Finnish parliamentary data and a semantic portal on top of it². The paper presents the vision and first results of the PARLIAMENTSAMPO extending our earlier papers on creating the knowledge graphs for the speeches [9] and MP networks [10] and a Finnish presentation on the project [11].

The paper first reviews related research on parliamentary data (Section 2). In Section 3, our vision of publishing and using Finnish parliamentary linked data on the SW is presented. After this, first results obtained in developing and using the PARLIAMENTSAMPO system in different ways are presented (Section 4). In conclusion, results of our work are summarized and using parliamentary data in research is considered on a more general level (Section 5).

2. Related Work on Parliamentary Data

Lots of parliamentary materials have been digitized in recent decades, arguably only second to newspapers [12]. For example, the Royal Library of Sweden has digitized Swedish printed parliamentary documents from 1521 to 1970. This collection³ is supplemented by the parliament’s own digital materials and, e.g., by the Westac research project⁴ at the Umeå University. Digitization has improved the accessibility and usability of parliamentary materials for both the public and the research community. Websites have been created that make it easy for users to browse and download materials. Examples include the website of the Lipad project⁵ [13] that digitized Canadian parliamentary materials, and the portal Italian House of Representatives⁶ that comprehensively presents the history of the Italian parliament in 1848–2018.

¹<https://www.w3.org/standards/semanticweb/>

²See the project homepage for more details, videos, and publications: <https://seco.cs.aalto.fi/projects/semparl/en/>.

³<http://data.riksdagen.se>

⁴<https://www.westac.se>

⁵<https://lipad.ca>

⁶<https://storia.camera.it>

Several parliamentary corpora have been formed from the minutes of the plenary debates, which make it possible to study the content of the speeches and their language; see, e.g., [14] and the CLARIN list of parliamentary corpora⁷. The TEI-based Parla-CLARIN scheme⁸ for session minutes has been developed within the CLARIN infrastructure, providing a common way to represent the corpora [15]. The related ParlaMint project⁹ brings together Parla-CLARIN-based national corpora. Parliamentary materials have also been transformed into the form of LD when creating the LinkedEP [3] system on the European Parliament's data, the Italian Parliament¹⁰, and the LinkedSaeima for the Latvian parliament [4].

The materials of the Parliament of Finland (PoF) have been digitized in various contexts but are difficult to use, as they have been produced separately from different periods and stored in different formats [9]. The usability of the materials is also hampered by their varying quality and lack of descriptive data [16]. Language corpora have been published on parliamentary debates, such as the Parliamentary Corpus of FIN-CLARIN's Language Bank¹¹ [17] which covers the years 2008--2016. It contains the speeches in a linguistically annotated form and also synchronized links to original plenary session videos [18]. The Voices of Democracy project has produced a research corpus that includes plenary minutes in 1980-2018 annotated grammatically as well as interviews of veteran MPs conducted by the PoF after 1988 [12]. The minutes of the parliamentary debates from 1991 to 2015 can also be found in the International Harvard Parlspeech Corpus [19], but we have identified gaps in the coverage in this corpus.

Digitized parliamentary materials offer a wide range of perspectives on different research topics and have been used in a variety of fields, such as linguistics, political science, media studies, economics, and history. The most important research material are the debates in the parliaments, through which one can study the language and its changes itself as well as the underlying societal phenomena at large [20]. Metadata makes it possible to structure the speeches, for example, between parties, gender, or professional groups. Blaxill and Beelen [21] have examined the content of women's parliamentary speeches, as well as the role of gender in the speeches of MPs in the British Parliament. Parliamentary debates have been used in thematic or conceptual analyses (cf., e.g., [22, 23, 24, 25, 26]) and to study the language and the opinions of the parties or MPs (e.g., [27, 28]). Parliamentary debates have been used in translation studies using, for example, the EuroParl Corpus¹² of the European Parliament debates.

The digitized material of the Finnish Parliament has been utilized to some extent in digital humanities and social scientific research. La Mela [16], also Kettunen and La Mela [26], have studied the history of the concept of Everyman's right, a Nordic right of public access to nature, with the digitized minutes of the Parliament, and examined their quality in PDF format. The digitized minutes have been utilized in the development of language technology methods, in this case the Finnish Semantic Tagger [26]. Similarly, Andrushchenko et al. [12] have used their grammatically structured corpus and a search tool to organize and analyze parliamentary debates in various research cases. Simola [29] has examined the differences in political speech between

⁷<https://www.clarin.eu/resource-families/parliamentary-corpora>

⁸<https://github.com/clarin-eric/parla-clarin>

⁹<https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora>

¹⁰<http://data.camera.it>

¹¹<http://korp.csc.fi>

¹²<https://www.statmt.org/europarl/>

parties throughout the parliamentary period 1907–2018, for which she compiled a separate research dataset combining the debates and the speaker data. Makkonen and Loukasmäki [30] have studied the plenary speeches given in Parliament of Finland in 1999–2014 and their content by using topic modeling. FIN-CLARIN’s Parliamentary Corpus has been used, for example, by Lillqvist et al. [31] in their study on debates about public debt. Previous search applications for Finnish parliamentary speech data are based mostly on traditional text search. Data analysis tools to examine the results are few, such as the concordance analysis of the Language Bank Korp, where the words found are visualized in their textual contexts and show some statistics of words occurrences in the search results. These applications cover only a small part of the entire time series of the Finnish parliamentary speeches.

3. ParliamentSampo Vision

The vision of the Semantic Parliament project [11] is to develop and implement in the living laboratory environment model shown in Fig. 1 for publishing and utilizing parliamentary materials as LOD on the SW. The work focuses on two core datasets:

1. **Minutes of Parliamentary Sessions** All Finnish parliamentary debates, totalling ca. 962 000 speeches and covering the existence of the PoF 1907–2021, have been transformed into a 1) Linked Data knowledge graph and into 2) Parla-CLARIN XML form. [9]
2. **Members of Parliament Data** A proposographical knowledge graph has been created for representing biographical data about all ca. 2800 Finnish MPs and other politicians during the same time period (1907–2021). [10]

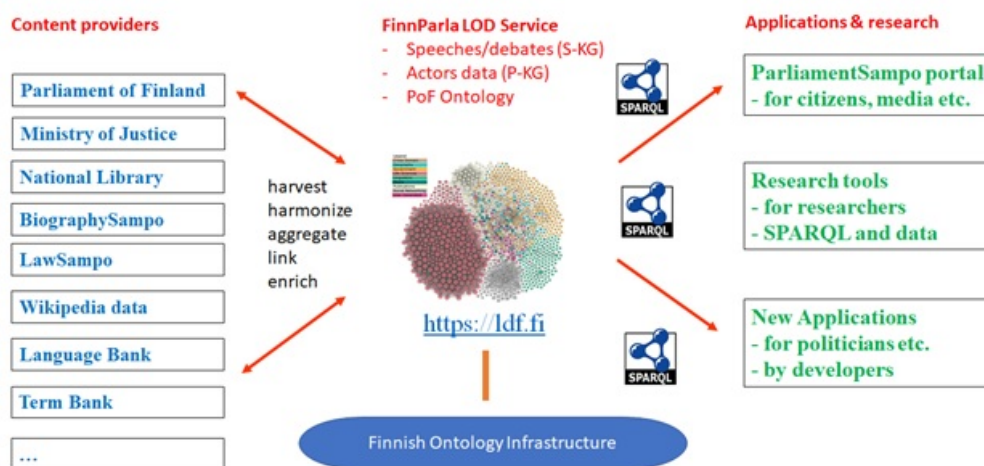


Figure 1: Vision: Linked Open Data publishing model of PARLIAMENTSAMPO

The left side of Fig. 1 shows content providers that produce data related to the PoF in their own local data silos, but in non-interoperable formats. For example, the LawSampo (LawSampo)

system publishes Finnish legislation, the results of parliamentary discussions, and case law data provided by the Ministry of Justice in Finland as the LOD service Semantic Finlex [32] and a semantic portal¹³ [33].

In the middle of Fig. 1, the data is aggregated, harmonized, enriched, interlinked, and published as a new FinnParla LOD service on the Linked Data Finland platform LDF.fi¹⁴ [34]. Its data model is based on 1) a new ontology describing the activities of the PoF and 2) a set of related vocabularies and ontologies describing, for example, (historical) places¹⁵, professions [35], people, and organizations. Notice that lots of additional documents of the PoF processes, such as propositions and bills, and votation data could be interlinked with the PARLIAMENTSAMPO system in the future using its open data infrastructure.

The right side of Fig. 1 depicts the ways of utilizing the FinParla data service as 1) a semantic portal, 2) using it in research by computational tools, and 3) for developing new applications.

The knowledge graph of the parliamentary speeches (S-KG) (cf. Fig. 1), contains speeches collected from all the minutes of the plenary sessions of the PoF since 1907. The S-KG was compiled from several initial formats: 1) From 1907 until the middle of 1999, the minutes are available only as scanned images embedded in PDF documents. This material was OCRed with minor manual corrections made. 2) From mid-1999 to the end of 2014, the material was available in HTML format at the Parliament's website¹⁶. 3) From the 2015 onwards, the minutes are available through the Finnish Parliament Open Data API¹⁷ in custom XML form. The data quality of S-KG has been deemed satisfactory, although there were issues related to OCR errors and the fact that there have been differences in how the transliteration and metadata of the minutes have been produced in the PoF. The data model of S-KG and the data transformation process are described in detail in [9].

The S-KG was interlinked to the MPs prosopographic knowledge graph P-KG (cf. Fig. 1). For example, speakers and the parties they represent are resources with URI identifiers described in the P-KG graph. The data publication about MPs is a knowledge graph (P-KG) covering all MPs who have worked in Finland [10]. At its core is an RDF conversion of XML-formatted data about MPs downloaded from the Open Data service¹⁸ of PoF. In addition to basic biographical information, such as times and places of birth and death, the data includes detailed information about the people's life events, such as studying, working life, political career, and publications written by the politicians.

The Finnish parliament's open data source has been supplemented and enriched with information extracted from the Finnish Government's website¹⁹ and Wikidata: in addition to MPs, some 200 other people with significant political history, such as presidents, ministers, and ombudsmen, have been added into the knowledge graph. For example, Mauno Koivisto has served as President and Prime Minister but never as an MP. The knowledge graph was also interlinked with the BiographySampo system [36], yet another example of the mutually interlinked "Sampo"

¹³LawSampo project: <http://seco.cs.aalto.fi/projects/lawlod/>

¹⁴Linked Data Finland service online: <https://ldf.fi/>

¹⁵<https://seco.cs.aalto.fi/projects/histoplaces/>

¹⁶<https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx>

¹⁷<https://avoindata.eduskunta.fi/#/fi/home>

¹⁸<https://avoindata.eduskunta.fi/#/fi/dbsearch>

¹⁹<https://valtioneuvosto.fi/hallitukset-ja-ministerit>

systems and LOD infrastructure²⁰ in use in Finland, that publishes biographies of ca. 13 600 significant Finnish persons as a LOD service and a semantic portal²¹ including biographies of 614 Finnish parliamentarians. The data model of P-KG, based on the CRM Bio extension [37] of CIDOC CRM²², is described in more detail in [10] including the transformation process of the data sources into RDF. The transformation and linking could be done fairly accurately as the primary data were already available in structured forms.

4. Using the PARLIAMENTSAMPO LOD Service

The goal of the PARLIAMENTSAMPO system is to provide the end users with flexible and rich possibilities for searching, browsing, and analyzing the PoF data. The new possibilities are offered by a standard SPARQL endpoint for 1) opening the data for external use, 2) for querying the endpoint and studying the results, 3) for data analysis using various tools and scripting, and 4) for developing new external applications, such as the PARLIAMENTSAMPO portal. These use cases are explored next in more detail with examples.

4.1. Exporting the Data for External Use

A simple way for a researcher to use PARLIAMENTSAMPO data is to download data from the data service for local use and then apply one's favourite tools for data analysis, such as spreadsheets, R²³ environment for statistical analysis, or Gephi²⁴ for network analysis. For filtering out subsets of interest in the big data, SPARQL querying can be used in flexible ways. It is also possible to install a local SPARQL server environment for linked data on one's own computer, for example Fuseki²⁵, which is also used in the LDF.fi service. The materials in the LDF.fi service are published using container technology (i.e., Docker²⁶), which means that installing the data, the server, and possible versioned software packages is automatic and effortless.

An example of using PARLIAMENTSAMPO data externally is reported in [20]. For this case study in political science, the Parla-CLARIN version was downloaded and a subset of the speeches 1960–2020 was filtered out and analyzed further using custom XML-based tools. The authors studied how the language used in discussing environmental politics has evolved in Finland in the speeches of different parties. Eleven central environmental terms were selected from a thesaurus²⁷ used by the PoF library, speeches where these terms were used were then extracted, and various quantitative analyses based on them were presented and compared with the strategy plans of the parties with qualitative interpretations. The analyses showed, for example, a constantly increasing intensity of environmental debates and a rhetorical shift of language from protecting the nature to issues of climate change.

²⁰LOD Infrastructure for Digital Humanities in Finland (LODI4DH): <https://seco.cs.aalto.fi/projects/lodi4dh/>

²¹BiographySampo portal is available at <https://biografiasampo.fi/>.

²²<https://cidoc-crm.org>

²³<https://www.r-project.org>

²⁴<https://gephi.org>

²⁵<https://jena.apache.org/documentation/fuseki2/>

²⁶<https://www.docker.com>

²⁷EKS Subject Headings: <https://www.eduskunta.fi/kirjasto/EKS/index.html?kieli=en>

4.2. Querying the Endpoint and Studying Results

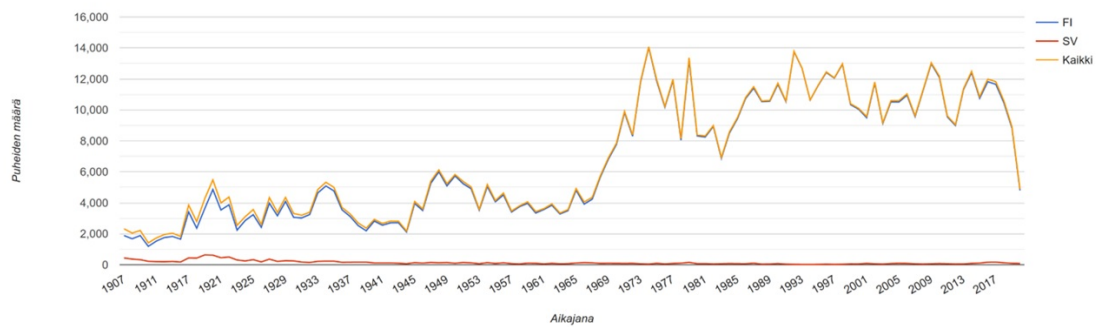


Figure 2: Number of speeches in different languages (y-axis) on the timeline (x-axis).

SPARQL is a flexible way to query RDF data. The search result is presented in a tabular format that can be examined as it is and be visualized and used for application-specific analyzes. For example, Fig. 2 shows a visualization of the number of speeches (y-axis) in the S-KG graph by language on a timeline from 1907 to 2021 (x-axis). Speeches in Finnish ('FI' in the figure) have clearly been given the most since the beginning ('Kaikki' in the figure denotes all the speeches). Originally, there have been more speeches in Swedish ('SV' in the figure) than today, but the number remains very small. The graphic was created using the YASGUI editor²⁸ [38], which can be used to edit SPARQL queries, target them to an online SPARQL endpoint, and to show the results using pre-implemented visualizations.

SPARQL is an expressive and flexible way to retrieve information from graphical data, and it is suitable for use by DH researchers. The SPARQL query used to generate Fig. 2 is shown below:

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> # For shortening URIs
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) # Variables in the result
           (count(?document-URI) as ?ALL ) WHERE {
  ?document-URI a semparls:Speech . # Graph pattern matched
  ?document-URI <http://purl.org/dc/terms/date> ?dateTime .
  BIND(STR(year(?dateTime)) as ?year)
  {
    BIND( <http://id.loc.gov/vocabulary/iso639-2/swe> as ?swe)
    ?document-URI dct:language ?swe .
  } UNION {
    BIND( <http://id.loc.gov/vocabulary/iso639-2/fin> as ?fin)
    ?document-URI dct:language ?fin .
  }
}
} GROUP BY ?year ORDER BY ASC(?year) # Grouping and ordering results yearly
```

This query above first introduces the namespaces used (PREFIX); they are used to make the URI references in the query syntactically shorter and simpler. In the next SELECT part of the

²⁸<https://yasgui.triply.cc>

query, all speeches and their languages are retrieved using a graph pattern formed by variables starting with `?`, which are fitted to the end point graph in all possible ways. The answer of the query is a table of all possible value assignments for the variables that make the query pattern to match the underlying data. The results are finally classified (GROUP BY) into groups according to language, sorted by year (ORDER BY), and finally it is summed up (COUNT) how many speeches there are in Finnish, Swedish, and in total. In the visualization, the variable `?year` forms the x-axis and the y-axis presents the annual number of speeches in different languages. When the speech graph was created, language recognition of speeches was done automatically. Typically this could be done accurately. However, sometimes OCR errors, for example, can make language recognition difficult, and therefore speeches whose language code could not be identified were excluded automatically from the query result.

4.3. Data-analysis by Scripting

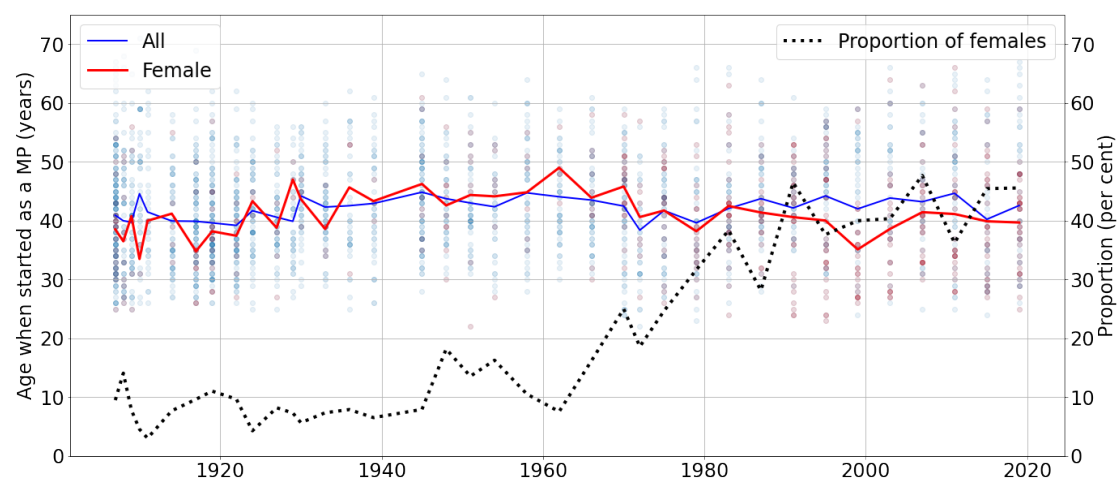


Figure 3: Annual starting age of new MPs and relative proportion of women.

The PoF data can be examined computationally, for example, using Python scripting and Jupyter notebooks in the Google Colab²⁹ environment. Then one can use the simple HTTP protocol to perform SPARQL queries and after this analyze and visualize query results using tools provided by the programming environment used, e.g., by Python libraries.

For example, Fig. 3 shows the ages of persons elected as MPs for the first time each year [10]. In the figure, the blue solid line shows the age of all MPs, and the age of women is shown in red. It can be seen from the graph that the starting age has remained almost constant throughout the parliamentary activities, but on the other hand, since 1980, women have been younger than men for some time when they started as MPs. The relative proportion of women in the PoF is shown by a black dotted line. Before the 1960s, the proportion remained at an average of

²⁹<https://colab.research.google.com>

10%, but has after this risen to 30–50%. The graphics in the image were implemented in Google Colab using standard Python libraries for data analysis.

Social Democratic Party of Finland	31	32	1	31	19	16	5	17	26	42	32	4	1	18	5	2	0	7	18	2	29	16
Centre Party	211	79	73	18	16	12	6	12	16	4	3	12	11	4	12	12	23	5	7	1	1	0
National Coalition Party	75	36	20	35	32	42	24	35	5	1	0	17	21	6	19	18	11	11	5	15	0	0
Swedish People's Party of Finland	42	15	6	6	13	17	22	2	6	2	0	7	4	10	2	8	1	7	2	6	0	0
Finnish People's Democratic League	10	11	0	5	4	1	2	2	0	6	15	1	0	4	0	0	0	2	5	2	8	13
Finnish Party	58	22	2	0	4	6	15	3	7	0	0	7	0	2	5	10	3	3	1	0	0	0
National Progressive Party	19	13	10	0	10	7	7	7	3	1	0	4	0	0	1	1	7	3	3	2	0	0
Young Finnish Party	22	11	8	0	6	3	7	2	9	1	0	6	0	0	1	1	4	4	3	0	0	0
Finns Party	0	0	0	6	2	3	0	0	0	0	0	0	15	3	2	0	0	4	0	3	0	1
Left Alliance	1	1	0	5	4	1	0	1	0	0	0	0	0	4	0	0	0	0	0	2	0	1
Green League	0	0	0	8	4	1	1	0	0	0	0	1	2	0	0	0	1	0	4	0	0	0
Finnish Rural Party	11	0	6	2	0	2	0	3	0	0	1	0	1	0	2	0	1	1	1	1	3	0
Liberals	0	1	0	4	4	3	3	3	0	1	0	1	0	3	1	0	0	1	0	2	0	0
Christian Democrats	0	0	0	3	2	2	1	2	0	0	0	0	0	1	3	0	0	0	0	0	0	0
	Farmer	Municipal councillor	Agronomist	Master of Social Sciences	Master of Arts	Lawyer with bench training	Professor	Managing director	Elementary school teacher	Reporter	Smallholder	Bank manager	Entrepreneur	Editor	Provost	Vicar	Agricultural councillor	Doctor in Philosophy	Editor-in-chief	Master of Science in Technology	Regional district secretary	Carpenter

Figure 4: Table of correlations between parties (y-axis) and MP professions (x-axis) [10]

Figure 4 shows a similarly formed tabular visualization of the correlation between the parties and the occupations of the MPs. Here only the most popular parties and occupations over the entire history of PoF are considered. The parties are presented in the horizontal rows of the table and the number of representatives of each profession is indicated in the vertical row corresponding to the occupation. The matrix shows, for example, that in the Centre Party, the National Coalition Party, and the Swedish People's Party the most common occupation is Farmer. On the other hand, Entrepreneur has been the most common occupations with the Finns Party.

The same visualization components can be reused in different contexts. For example, the matrix visualization of Fig. 4 is re-used in Fig. 5 for analyzing interruptions of speeches of the current PoF. The y-axis lists the most active speakers and x-axis the MPs that have interrupted their speeches. For example, of the interrupted speeches of MP Annika Saarikko (Centre Party), the current Minister of Finance, 46% are due to MP Ben Ben Zyskowics, representing the National Coalition Party in opposition, and 18% to MP Jukka Gustafsson representing the party SDP in the government, indicating possibly different opinions inside the government.

Puhuja	Heinonen, Timo	17%	0%	26%	11%	2%	14%	11%	7%	5%	5%	1%	2%
	Kiljunen, Kimmo	7%	21%	0%	13%	2%	0%	2%	12%	18%	5%	15%	5%
	Kiuru, Krista	46%	11%	5%	1%	25%	1%	2%	1%	0%	2%	3%	3%
	Marin, Sanna	39%	12%	11%	4%	16%	1%	5%	1%	2%	4%	4%	1%
	Orpo, Petteri	11%	19%	37%	2%	5%	2%	15%	0%	0%	2%	2%	3%
	Lindtman, Antti	30%	22%	5%	10%	14%	4%	2%	2%	3%	5%	2%	2%
	Vanhanen, Matti	26%	6%	2%	24%	12%	1%	4%	2%	6%	7%	6%	2%
	Viitanen, Pia	20%	42%	2%	2%	6%	4%	2%	6%	7%	0%	2%	6%
	Lintilä, Mika	36%	10%	9%	14%	7%	0%	11%	0%	2%	1%	7%	2%
	Saarikko, Annika	46%	9%	4%	4%	7%	0%	18%	1%	1%	5%	6%	0%
	Zyskowitz, Ben	0%	15%	10%	25%	7%	15%	7%	1%	3%	1%	6%	7%
	Hoskonen, Hannu	2%	25%	3%	30%	2%	3%	0%	5%	20%	5%	5%	0%
		Zyskowitz, Ben											
	Heinonen, Timo												
	Arhinmäki, Paavo												
	Meri, Leena												
	Sarkomaa, Sari												
	Kiljunen, Kimmo												
	Gustafsson, Jukka												
	Eerola, Juho												
	Kankaanniemi, Toimi												
	Halla-aho, Jussi												
	Niikko, Mika												
	Vehviläinen, Anu												
	Keskeyttäjä												

Figure 5: Table of correlations that indicate how the most active speakers (y-axis) of the current PoF have been interrupted by other MPs (x-axis).

4.4. Using the PARLIAMENTSAMPO Portal

The PARLIAMENTSAMPO portal, based on the Sampo model [39] and the Sampo-UI framework [40], demonstrates how the FinnParla data service can be used for developing applications for DH research. In the portal, the data can be filtered using faceted search [41] based on ontologies, and the results can then be analyzed with the help of seamlessly integrated visualization and data analysis tools. The data can be accessed along two application views for studying 1) speeches and 2) MPs. For example, in Fig. 6, the user has selected the Plenary Speeches view, which shows the search facets Content, Speaker, Party, (Speech) Type, Language, and Date on the left. The search result, i.e., the speeches found, is shown by default in tabular form on the right. The user has written a query “suomettum*” in the Content text facet, in which case only speeches that contain the word “suomettuminen” (Finlandization) in its various inflectional forms have been filtered into the search result, as the wildcard “*” matches any string. The user has also limited the result on the Date facet to speeches given since June 4, 1945, when Parliament began to convene after the World War II. The result in this case is 177 speeches, shown in a table (with paging). By selecting the tab “Timeline”, the yearly amount of speeches is visualized as a function of time.

In faceted search, the filtering selections can be made flexibly in any order, and the search engine calculates a hit count for each subsequent facet selection, which tells how many results would be obtained in the result set if the selection in question is made next. For example, in

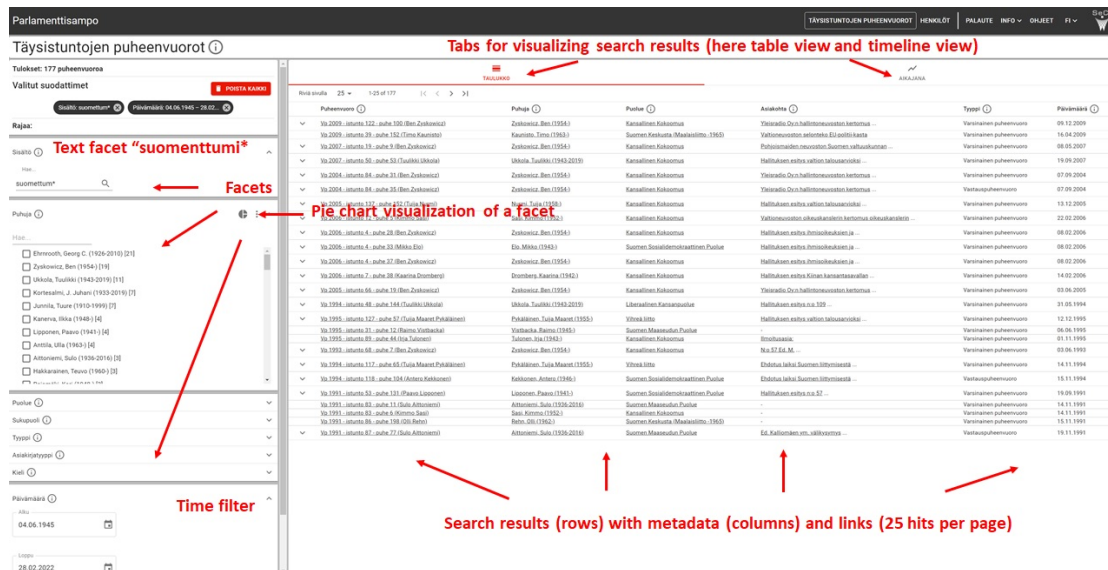


Figure 6: Using faceted search to filter out speeches of interest.

the Speaker facet, a click on “Junnila, Tuure (1910-1999) [7]” selects MP Tuure Junnila’s seven speeches that mention “Finlandization”. The selection facets are created automatically using the parliamentary ontology and knowledge graphs of the FinParla data. The hit count allows the user to be directed to selections that do not lead to dead ends where the result set is empty. In addition, the hit numbers provide an opportunity to investigate the result set statistically along different facet dimensions. For example, a click on the pie symbol of the Speaker facet opens the pie chart of Fig 7 which shows how many different speakers mention “Finlandization” in their speeches. The most active MPS in this case are Mr. Georg C. Ehrnrooth (21 speeches) and Mr. Ben Zyskowicz (19 speeches), two active right-wing politicians concerned with the concept.

In accordance with the Sampo model, a number of pre-implemented data analysis tools and visualizations, similar to those shown in the figures above, can be integrated into the application perspectives of the PARLIAMENTSAMPO portal. In the future, the tools and visualizations can be found alongside the table visualization in Fig. 6 on their own tabs in the same way as, for example, in the AcademySampo’s user interface [42]; the components of the Sampo-UI framework [40] are reused in the implementation of both portals. Through these tools and visualizations, the project explores the potential of Artificial Intelligence for knowledge discovery in DH research [43], i.e., how could PARLIAMENTSAMPO assist a researcher in finding research problems, in solving them, and also in explaining solutions?

5. Discussion

In the context of political research, the parliamentary speech is considered an important form of political communication and political struggle. A parliamentary speech is not just any speech,

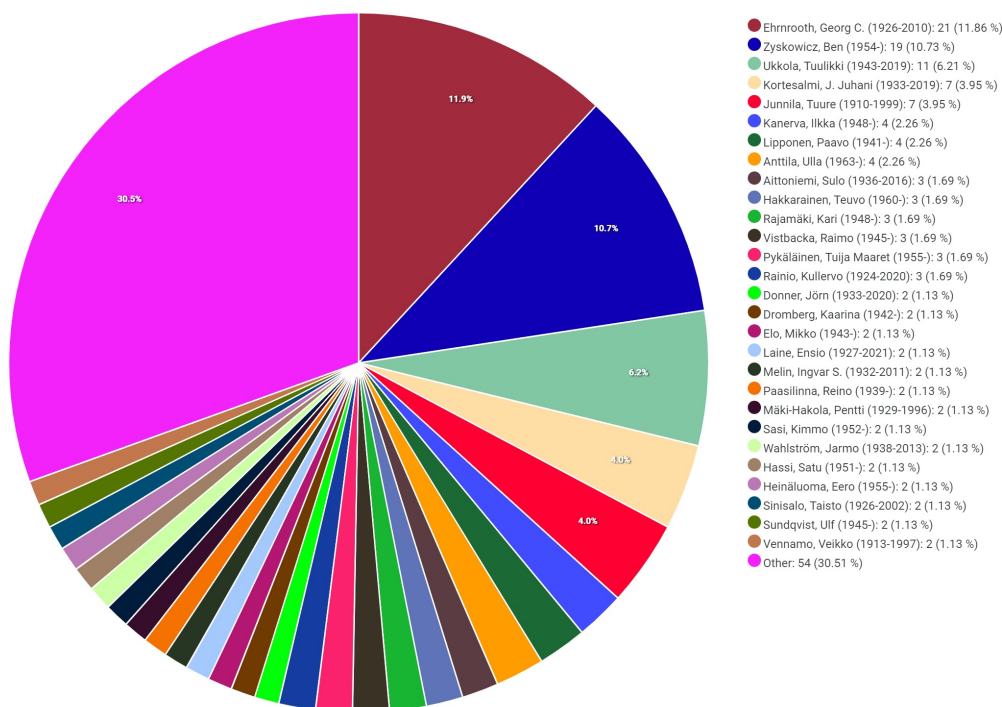


Figure 7: Speeches containing the word "suomettuminen" in any form as a pie chart, calculated according to the distribution of MPs on the Speaker facet. The speakers are listed on the right.

but has its own structure and its own rules, which at the same time reflect the general position of the parliament. In addition, a parliamentary speech is an instrument of political struggle to expose competing goals, challenge the views of an adversary, and unlock deadlocked settings. Thus, a speech in a parliament is always also a political act, in which the words used are the weapons of political decision-making and which not only tell about the issues under discussion, but also reveal the different positions, values, and points of view of the speakers. [44]

Traditionally, parliamentary speeches have been studied by close reading and using content analysis, discourse analysis, and various methods of rhetorical research. However, digitalization has also entered this traditional area of research more and more, as data on parliamentary debates in various countries have become increasingly available in the form of open data. In the case of the PoF, the digitization of parliamentary documents has progressed at a reasonable speed, and some of the material has also been available through the parliament's open data service. The availability of the data and also the quality of the available data has improved in recent years, but there are still significant differences with similar data in different countries.

The work on PARLIAMENTSAMPO is an important step in utilizing the plenary debates in PoF as part of the field of humanities research. Although the materials have always been available to researchers manually and for some years also electronically digitized in PDF format, the

machine-readable data corpus now being prepared and published as a data service, together with the PARLIAMENTSAMPO portal, will integrate parliamentary plenary debates and other open materials into the DH and national information infrastructure. This means in practice, for example, the opportunity for political scientists, historians, and linguists to extract, model, analyze, and visualize parliamentary speech through exploratory research, using a vast body of data covering the entire period of the modern PoF since 1907.

The possibility of exploratory data analysis opens up completely new possibilities and perspectives for the study of parliamentary speech. In traditional close reading, the researcher is forced to delimit the material strongly already at the collection stage, which usually happens through either temporal or thematic delimitation – that is, either by focusing on a limited time period or on limited themes. Digital methods make it possible to study the material without such limitations, and thus to examine it, for example, with fully automatic or semi-automatic classification methods. In this way, it may be possible to find, for example, new themes and topics that have been sidelined in research in the past (cf., e.g., [45, 46]).

On the other hand, distant reading and classification of data without strong presuppositions also allows for a critical examination of previous research results, when the themes/topics generated by distant reading can be compared with the results obtained by other methods [47].

Another example of the possibilities offered by data is research on the language of politics and its long-term change (e.g., [48, 49, 50, 51, 52, 53]). Parliamentary big data enables large-scale and systematic application of language technology methods. Although parliamentary speech is also linguistically its own special form of speech, parliamentary speech also lives in time and thus reflects both the wider linguistic development and the social atmosphere of discussion and word choices that occur in it [30]. At the same time, the extensive data offer an opportunity to study the change in language use, for example, whether the social debate climate is polarized or “brutalized”, as politicians and media actors have repeatedly suggested in recent years.

The third opportunity offered by parliamentary data relates to linking the use of language more broadly to other social contexts of language users, such as education, age, and social networks. Language can also be approached in policy research on the assumption that language always reflects the wider world of values and ideas of its user, as well as his or her social status and context. Discursive coalitions, which can be constructed based on the language use of the speakers, thus offer an interesting opportunity to detach oneself from the frame of reference set by, for example, the party background and to focus analytical attention on networks built through the use of language. In previous studies, this type of approach has been able to connect experts to different ideological positions by analyzing the content of their texts [54], which we think can be well applied to the classification of MPs.

A few examples have been highlighted above where the utilization of parliamentary data would seem to allow for significant new research openings in parliamentary research. However, in the spirit of exploratory data analysis, it is worth highlighting the as-yet-unknown possibilities that gradually emerge as researchers begin to outline new hypotheses and research questions by examining and analyzing data. The potential of large datasets is surprising in their potential, which on the one hand requires an open-minded attitude towards the data and on the other hand underscores the growing responsibility of researchers working in data analysis. When it is no longer possible for the researcher to know the material he or she is using thoroughly, he or she must know the phenomena that are the subject of the material thoroughly. Only in this

way is it possible to assess which findings transmitted through excavation, analysis, modeling, or visualization are truly relevant.

Acknowledgements Our work is funded by the Academy of Finland and is also related to the EU project InTaVia³⁰ and the EU COST action Nexus Linguarum³¹. The project uses the computing resources of the CSC – IT Center for Science.

References

- [1] M. Hidén, H. Honka-Hallila, *Miten eduskunta toimii (How Parliament of Finland works)*, Edita Publishing, Helsinki, 2006.
- [2] C. Benoît, O. Rozenberg (Eds.), *Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures*, Edward Elgar Publishing, 2020. doi:10.4337/9781789906516.
- [3] A. Van Aggelen, L. Hollink, M. Kemman, M. Kleppe, H. Beunders, The debates of the European Parliament as Linked Open Data, *Semantic Web – Interoperability, Usability, Applicability* 8 (2017) 271–281. doi:10.1007/s42001-019-00060-w.
- [4] U. Bojārs, R. Dargis, U. Lavrinovičs, P. Paikens, Linkedsaeima: A linked open dataset of Latvia’s parliamentary debates, in: *Semantic Systems. The Power of AI and Knowledge Graphs. SEMANTiCS 2019*, Springer, 2019, pp. 50–56. doi:10.1007/978-3-030-33220-4_4.
- [5] T. Heath, C. Bizer, *Linked Data: Evolving the Web into a Global Data Space (1st edition)*, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, 2011. URL: <http://linkeddatabook.com/editions/1.0/>.
- [6] E. Hyvönen, *Publishing and Using Cultural Heritage Linked Data on the Semantic Web*, Synthesis Lectures on the Semantic Web: Theory and Technology, Morgan & Claypool, Palo Alto, CA, USA, 2012.
- [7] S. Staab, R. Studer (Eds.), *Handbook on Ontologies (2nd Edition)*, Springer, 2009.
- [8] P. Hitzler, M. Krötzsch, S. Rudolph, *Foundations of Semantic Web technologies*, Springer, 2010.
- [9] L. Sinikallio, S. Drobac, M. Tamper, R. Leal, M. Koho, J. Tuominen, M. L. Mela, E. Hyvönen, Plenary debates of the parliament of finland as linked open data and in parla-clarin markup, in: *3rd Conference on Language, Data and Knowledge, LDK 2021*, Schloss Dagstuhl- Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, 2021, pp. 1–17. URL: <https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASiCS-LDK-2021-8.pdf>.
- [10] P. Leskinen, E. Hyvönen, J. Tuominen, Members of Parliament in Finland knowledge graph and its linked open data service, in: *of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, 2021*, pp. 255–269. URL: <https://ebooks.iospress.nl/volumearticle/57420>. doi:10.3233/SSW210049.
- [11] E. Hyvönen, L. Sinikallio, P. Leskinen, S. Drobac, J. Tuominen, K. Elo, M. L. Mela, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, *Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet*, Informaatiotutkimus 40 (2021). URL: <https://doi.org/10.23978/inf.107899>.

³⁰<https://intavia.eu>

³¹<https://nexuslinguarum.eu>

- [12] M. Andrushchenko, K. Sandberg, R. Turunen, J. Marjanen, M. Hatavara, J. Kurunmäki, T. Nummenmaa, M. Hyvärinen, K. Teräs, J. Peltonen, J. Nummenmaa, Using parsed and annotated corpora to analyze parliamentarians' talk in Finland, *Journal of the Association for Information Science and Technology* 185 (2021) 1–15. doi:10.1002/asi.24500.
- [13] K. Beelen, T. A. Thijm, C. Cochrane, K. Halvemaan, G. Hirst, M. Kimmins, S. Lijbrink, M. Marx, N. Naderi, L. Rheault, R. Polyanovsky, T. Whyte, Digitization of the Canadian parliamentary debates, *Canadian Journal of Political Science* 50 (2017) 849–864. doi:10.1017/S0008423916001165.
- [14] E. Lapponi, M. G. Søyland, E. Velldal, S. Oepen, The talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016, *Lang Resources & Evaluation* 52 (2018) 873–893. doi:10.1007/s10579-018-9411-5.
- [15] A. Pancur, T. Erjavec, The siParl corpus of Slovene parliamentary proceedings, in: *Proceedings of the Second ParlaCLARIN Workshop*, European Language Resources Association, 2020, pp. 28–34. URL: <https://www.aclweb.org/anthology/2020.509parlaclarin-1.6>.
- [16] M. La Mela, Tracing the emergence of nordic allemansrätten through digitised parliamentary sources, in: M. Fridlund, M., Oiva, P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history*, Helsinki University Press, 2020, pp. 181–197. doi:10.33134/HUP-5-11.
- [17] M. Lennes, FIN-CLARIN and language bank parliamentary data workshop “digital parliamentary data and research”, 2019. URL: <https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/workshop-digital-parliamentary-data-and-research>.
- [18] A. Mansikkaniemi, P. Smit, M. Kurimo, Automatic construction of the Finnish parliament speech corpus, in: *Proc. Interspeech 2017*, 2017, pp. 3762–3766. doi:10.21437/Interspeech.2017-1115.
- [19] C. Rauh, P. De Wilde, J. Schwalbach, The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states (V1), 2017. doi:10.7910/DVN/E4RSP9.
- [20] K. Elo, J. Karimäki, Luonnonsuojelusta ilmastopoliittikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020, *Politiikka* 63 (2021). URL: <https://journal.fi/politiikka/article/view/109690>. doi:10.37452/politiikka.109690.
- [21] L. Blaxill, K. Beelen, A feminized language of democracy? the representation of women at Westminster since 1945, *Twentieth Century British History* 27 (2016) 412–449. doi:10.1093/tcbh/hww028.
- [22] K. Quinn, B. Monroe, M. Colaresi, M. H. Crespin, D. R. Radev, How to analyze political attention with minimal assumptions and costs, *American Journal of Political Science* 54 (2010) 209–228. doi:10.1111/j.1540-5907.2009.00427.x.
- [23] H. Baker, B. V., M. T., Digitization of the Canadian parliamentary debates, in: T. Säily, A. Nurmi, M. Palander-Collin, A. Auer (Eds.), *Exploring future paths for historical sociolinguistics*, John Benjamins, Amsterdam, 2017, pp. 83–107. doi:10.1017/S0008423916001165.
- [24] J. Guldi, Parliament's debates about infrastructure: An exercise in using dynamic topic models to synthesize historical change, *Technology and Culture* 60 (2019) 1–33. doi:10.1353/tech.2019.0000.

- [25] P. Ihalainen, A. Sahala, Evolving conceptualisations of internationalism in the UK parliament: Collocation analyses from the League to Brexit, in: M. Fridlund, M., Oiva, P. Paju (Eds.), *Digital histories: Emergent approaches within the new digital history*, Helsinki University Press, 2020, pp. 199–219. doi:10.33134/HUP-5-12.
- [26] K. Kettunen, M. La Mela, Semantic tagging and the nordic tradition of everyman’s rights, *Digital Scholarship in the Humanities* (2021). doi:10.1093/llc/fqab052.
- [27] G. Abercrombie, R. Batista-Navarro, Sentiment and position-taking analysis of parliamentary debates: a systematic literature review, *Journal of Computational Social Science* 3 (2012) 245–270. doi:10.1007/s42001-019-00060-w.
- [28] M. Magnusson, R. Öhrvall, K. Barrling, D. Mimno, Voices from the far right: a text analysis of Swedish parliamentary debates, *SocArXiv* (2018). doi:10.31235/osf.io/jdsqc.
- [29] S. Simola, A century of partisanship in Finnish political speech, 2020. URL: <https://sites.google.com/site/sallasimolaecon/home/research>.
- [30] K. Makkonen, P. Loukasmäki, Eduskunnan täysistunnon puheenaiheet 1999–2014: Miten käsitellä LDA-aihemalleja?, *Politiikka* 61 (2019) 127–159. URL: <https://journal.fi/politiikka/article/view/77163>.
- [31] E. Lillqvist, I. K. Kavonius, M. Pantzar, “velkakello tikittää”: Julkisyhteisöjen velka suomalaisessa mielikuvastossa ja tilastoissa 2000–2020, *Kansantaloudellinen Aikakauskirja* 116 (2020) 581–607. URL: <https://journal.fi/politiikka/article/view/77163>.
- [32] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen, E. Hyvönen, Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web, in: *Knowledge of the Law in the Big Data Age*, volume 317 of *Frontiers in Artificial Intelligence and Applications*, IOS Press, 2019, pp. 212–228.
- [33] E. Hyvönen, M. Tamper, E. Ikkala, S. Sarsa, A. Oksanen, J. Tuominen, A. Hietanen, Publishing and using legislation and case law as linked open data on the semantic web, in: *The Semantic Web: ESWC 2020 Satellite Events*, Springer, 2020, pp. 110–114. doi:10.1007/978-3-030-62327-2_19.
- [34] E. Hyvönen, J. Tuominen, M. Alonen, E. Mäkelä, Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets, in: *The Semantic Web: ESWC 2014 Satellite Events, Revised Selected Papers*, Springer-Verlag, 2014, pp. 226–230. URL: https://doi.org/10.1007/978-3-319-11955-7_24.
- [35] M. Koho, L. Gasbarra, J. Tuominen, H. Rantala, I. Jokipii, E. Hyvönen, AMMO Ontology of Finnish Historical Occupations, in: *Proceedings of the The First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH’19)*, volume 2375, CEUR Workshop Proceedings, 2019, pp. 91–96. URL: <http://ceur-ws.org/Vol-2375/>.
- [36] E. Hyvönen, P. Leskinen, M. Tamper, H. Rantala, E. Ikkala, J. Tuominen, K. Keravuori, BiographySampo – publishing and enriching biographies on the semantic web for digital humanities research, in: *The Semantic Web. 16th International Conference, ESWC 2019, Proceedings*, Springer, 2019, pp. 574–589. doi:10.1007/978-3-030-21348-0.
- [37] J. Tuominen, E. Hyvönen, P. Leskinen, io CRM: A data model for representing biographical data for prosopographical research, in: *Proceedings of the Second Conference on Biographical Data in a Digital World 2017 (BD2017)*, volume 2119, CEUR Workshop Proceedings, 2018, pp. 59–66. URL: <http://ceur-ws.org/Vol-2119/paper10.pdf>.
- [38] L. Rietveld, R. Hoekstra, The YASGUI family of SPARQL clients, *Semantic Web – Interop-*

- erability, Usability, Applicability 8 (2017) 373–383. doi:10.3233/SW-150197.
- [39] E. Hyvönen, Digital humanities on the Semantic Web: Sampo model and portal series, *Semantic Web – Interoperability, Usability, Applicability* (2022). Accepted, <https://seco.cs.aalto.fi/publications/2021/hyvonen-sampo-model-2021.pdf>.
- [40] E. Ikkala, E. Hyvönen, H. Rantala, M. Koho, Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces, *Semantic Web – Interoperability, Usability, Applicability* 13 (2022) 69–84. doi:10.3233/SW-210428.
- [41] Y. Tzitzikas, N. Manolis, P. Papadakos, Faceted exploration of RDF/S datasets: a survey, *Journal of Intelligent Information Systems* 48 (2017) 329–364.
- [42] E. Hyvönen, P. Leskinen, H. Rantala, E. Ikkala, J. Tuominen, Akatemiasampo-portaali ja -datapalvelu henkilöiden ja henkilöryhmien historialliseen tutkimukseen, *Informaatio-tutkimus* 40 (2021) 28–56. URL: <https://journal.fi/inf/article/view/102656/64169>.
- [43] E. Hyvönen, Using the Semantic Web in Digital Humanities: Shift from Data Publishing to Data-analysis and Serendipitous Knowledge Discovery, *Semantic Web – Interoperability, Usability, Applicability* 11 (2020) 187–193. doi:10.3233/SW-190386.
- [44] K. Palonen, Eduskunnasta puhekunnaksi? Parlamentarismi retorisenä politiikkana, *Politiikka* 47 (2005) 141–148.
- [45] D. Mimno, Topic Regression, Ph.D. thesis, University of Massachusetts Amherst, 2012. URL: https://scholarworks.umass.edu/open_access_dissertations/520.
- [46] T. R. Tangherlini, P. Leonard, Trawling in the sea of the great unread: Sub-corpus topic modeling and humanities research, *Poetics* 41 (2013) 725–749. doi:10.1016/j.poetic.2013.08.002.
- [47] T. Ylä-Anttila, V. Eranti, Aihemallinnuksesta kehysmallinnukseen, *Politiikka* 60 (2005) 148–156. URL: <http://elektra.helsinki.fi/se/p/politiikka/60/2/aihemall.pdf>.
- [48] P. DiMaggio, M. Nag, D. Blei, Exploiting affinities between topic modeling and the sociological perspective on culture: Application to newspaper coverage of U.S. Government arts funding, *Poetics* 41 (2013) 570–606. doi:10.1016/j.poetic.2013.08.004.
- [49] C. Jacobi, W. van Atteveldt, K. Welbers, Quantitative analysis of large amounts of journalistic texts using topic modelling, *Poetics* 4 (2016) 89–106. doi:10.1080/21670811.2015.1093271.
- [50] S. Purhonen, A. Toikka, “Big Datan” haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät: esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uuden vuodenpuheista 1935–2015, *Sosiologia* 53 (2016) 6–27. URL: <http://elektra.helsinki.fi/se/s/0038-1640/53/1/bigdatan.pdf>.
- [51] S.-M. Laaksonen, M. Nelimarkka, Omat ja muiden aiheet: Laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta, *Politiikka* 60 (2018) 132–147.
- [52] A. Törnberg, P. Törnberg, Muslims in social media discourse: Combining topic modeling and critical discourse analysis, *Discourse, Context and Media* 13 (2016) 132–142. doi:10.1016/j.dcm.2016.04.003.
- [53] J. B. Mountford, Topic modeling the red pill, *Social Sciences* 7 (2018). doi:10.3390/socsci7030042.
- [54] Z. Jelveh, B. Kogut, S. Naidu, Detecting latent ideology in expert text: Evidence from academic papers in economics, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), ACL, 2018*, pp. 1804–1809.