

# ParliamentSampo Infrastructure for Publishing the Plenary Speeches and Networks of Politicians of the Parliament of Finland as Open Data Services

Eero Hyvönen<sup>1,2</sup>, Petri Leskinen<sup>1,2</sup>, Laura Sinikallio<sup>2,1</sup>, Senka Drobac<sup>2,1</sup>, Rafael Leal<sup>1,2</sup>, Matti La Mela<sup>2,3</sup>, Jouni Tuominen<sup>2,1</sup>, Henna Poikkimäki<sup>1</sup> and Heikki Rantala<sup>1</sup>

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

<sup>2</sup>*Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland*

<sup>3</sup>*Department of ALM, Uppsala University, Sweden*

## Abstract

This paper presents a new open infrastructure called PARLIAMENTSAMPO for studying the parliamentary culture, language, and activities of politicians in Finland. For the first time, the entire time series of some million plenary speeches of the Parliament of Finland (PoF) have been converted into data and data services in unified formats, including CSV, Parla-CLARIN, ParlaMint, and RDF Linked Open Data (LOD). The speech data have been interlinked with a knowledge graph about the activities of the Members of Parliament (MP) and other speakers in the plenary sessions of the PoF, enriched by data linking from external data sources into a broader ontology-based LOD service. Knowledge extraction techniques based on Natural Language Processing (NLP) were used for automatic semantic annotations and topical classification of the speeches. The data and data services have been used in Digital Humanities (DH) research projects and for application development, especially for developing the in-use semantic portal PARLIAMENTSAMPO. The infrastructure is openly available on the Web using the CC BY 4.0 license.

## Keywords

parliamentary studies, semantic portals, linked data, digital humanities


*Paper presented at the publication event of the ParliamentSampo infrastructure, University of Helsinki, February 14th, 2023*

## 1. Introduction

Parliaments enact new laws, oversee the work of the government, and decide on the state budget. The most prominent part of the work of parliaments are the public plenary sessions, in which the Members of Parliament (MP) discuss and vote on issues that arise. Openness and transparency of this work is important for the voters, media, researchers, and the parliamentarians themselves. The parliaments in different countries therefore publish minutes of plenary sessions and related documents openly to the public. Parliamentary data is used in many areas of research: they provide a wealth of information on the state and functioning of democratic systems, political life, language, and culture [1].

---

 [eero.hyvonen@aalto.fi](mailto:eero.hyvonen@aalto.fi) (E. Hyvönen)

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



© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

This paper presents a data publishing infrastructure called PARLIAMENTSAMPO about the speeches and politicians of the Parliament of Finland (PoF), starting from 1907 when PoF was established. To cater different user needs, the data is published in different formats, including CSV tables, XML-based formats Parla-CLARIN and ParlaMint, and as Linked Open Data knowledge graphs in RDF form. The usability of the infrastructure has been tested in Digital Humanities (DH) research projects and in developing a semantic portal PARLIAMENTSAMPO in use on top of the LOD service SPARQL endpoint. This paper extends our earlier papers about PARLIAMENTSAMPO [2, 3, 4, 5, 6] by focusing on the data resources and services available on the Web that constitute the new, openly available infrastructure published on February 14th, 2023<sup>1</sup>.

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 Semantic Web is presented. After this the data production pipelines of PARLIAMENTSAMPO and data models of their different outputs are explained (Sections 4 and 5). Examples of using the PARLIAMENTSAMPO data in different ways are given to illustrate the usability of the infrastructure in research (Section 6). In conclusion, results of our work are summarized (Section 7).

## 2. Related Work on Publishing and Using Parliamentary Data

In recent years, parliamentary debate corpora and digital parliamentary datasets have been created from the documents of both historical and contemporary parliaments [7, 8]. This digitization work has been conducted by the parliaments themselves, but also as part of research projects and by cultural heritage institutions. The aim has been to improve the accessibility and usability of these key documents of democratic societies for the public, but at the same time, the digitization has allowed researchers to engage in novel and interdisciplinary research using the new parliamentary data [7, 8]. Moreover, as part of the digitization and the research initiatives, web user interfaces and data services have been developed that allow to browse, study, and download the digitised materials.<sup>2</sup>

Among the recent parliamentary data publications, the projects have focused on the curation, annotation, and harmonization of the national parliamentary corpora, and also applied semantic web technologies for linking and enriching the parliamentary data with other datasets. In the pioneering Linked Data of the European Parliament (LinkedEP), the debates of the European Parliament and the political affiliation information were connected as linked data into other datasets such as DBpedia and the EuroVoc thesaurus [10]. Moreover, the LinkedEP data was made available through a SPARQL endpoint and an online user interface. Other examples of linked data parliament initiatives are the LinkedSaeima for the Latvian parliament [11], the Italian Parliament data<sup>3</sup>, and the historical Imperial Diet of Regensburg of 1576 project [12]. A key initiative for harmonization and annotation of national parliamentary corpora is the ParlaMint project part of the CLARIN infrastructure.<sup>4</sup> The ParlaMint project applies the TEI-based Parla-CLARIN scheme<sup>5</sup>, and aims to create uniformly annotated multilingual parliamentary corpora

---

<sup>1</sup>Publication event homepage: <https://seco.cs.aalto.fi/events/2023/2023-02-14-parlamenttisampo/>

<sup>2</sup>See e.g. the Lipad project and the Canadian Hansard, <https://lipad.ca> [9]

<sup>3</sup><http://data.camera.it>

<sup>4</sup><https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora>

<sup>5</sup><https://github.com/clarin-eric/parla-clarin>

with its partners. The current ParlaMint II involves 27 national parliamentary corpora [13] (see also [14]).

The minutes of the Parliament of Finland (PoF) have been digitized by the Parliament itself, but are challenging to use, as they have been produced separately in and from different periods, stored in different data formats, vary in quality, and lack descriptive data [2, 15]. Finnish parliamentary debates have been published as language corpora, for example by the FIN-CLARIN's Language Bank<sup>6</sup> [16], where the Parliamentary corpus contains linguistically annotated plenary debates and also links to the session videos [17]. The Voices of Democracy project has produced a research corpus that includes grammatically annotated plenary minutes in 1980–2018 as well as interviews of veteran MPs conducted by the PoF after 1988 [18]. The speeches of the Finnish parliamentar from 1991 to 2015 have been included also in the International Harvard Parlspeech Corpus [19], but which has gaps in the coverage.

Digitized parliamentary documents are used in many fields, such as linguistics, political science, legal studies, media studies, economics, and history. The main material used in research are the parliamentary debates combined with the political affiliation information, which allow to study, among others, (political) language and its use, legislative processes and political decision-making, and the debated societal issues (see for example [7, 8]). Metadata and annotation makes it possible to structure the speeches, for example, between parties, gender, government-opposition role, or professional groups, and to filter and analyse the speeches based on the annotated features. Moreover, the parliamentary data allow long-term studies as the data often extends over several decades or even a century [20]. Parliamentary debates have been used in thematic or conceptual analyses (cf., e.g., [21, 22, 20, 23, 24, 25]) and to study the language and the opinions of the parties or MPs (e.g., [26, 27, 18, 28, 29]). Parliamentary debates have been used in translation studies using, for example, the EuroParl Corpus<sup>7</sup> of the European Parliament debates.

The debates of the PoF have been employed previously in several social scientific and linguistic studies. La Mela [15], also Kettunen and La Mela [24], have studied the history of Nordic right of public access to nature, and examined the quality of the previous PoF open data. The digitized minutes have been utilized in the development of language technology methods[24]. Andrushschenko et al. [18] have used their grammatically structured corpus for selected digital humanities research cases. Simola [30] has explored the differences in political speech between parties in the long term (1907–2018), and Makkonen and Loukasmäki [31] have used topic modeling to study the plenary debates of PoF in 1999–2014. FIN-CLARIN's Parliamentary Corpus has been used, for example, by Lillqvist et al. [32] in their study on debates about public debt. Previous applications for Finnish parliamentary data cover only a small part of the entire time series of the Finnish parliamentary speeches. 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 with statistics about word occurrences.

---

<sup>6</sup><http://korp.csc.fi>

<sup>7</sup><https://www.statmt.org/europarl/>

### 3. ParliamentSampo Vision

The section outlines the vision and motivations behind building the PARLIAMENTSAMPO infrastructure [33].

#### 3.1. Parliamentary Debates as FAIR data for Problem Solving

The minutes of the plenary session of PoF have been available as printed books at the library of the PoF and now also through the PoF's open data service as scanned PDF documents, HTML pages or XML format, depending on which parliamentary sessions are in question<sup>8</sup>. However, they have not been published as data in accordance with modern FAIR principles in a Findable and Accessible (Interoperable) and Re-usable format for searching, browsing, and data analytic applications<sup>9</sup>. If the user knows during which parliament a speech was given, he could download, e.g., a scanned minutes book, which can be over thousand pages long, and search for the speech and other information in the document. But if one wants, for example, to find out the answers to the following questions, the this kind of online service and research method based on downloading and close-reading documents is not a viable solution:

1. **Question:** Which MP was the first to speak about “NATO” in the PoF? **Answer:** Mr. Yrjö Enne, SKDL party, 27 May 1959
2. **Question:** Who and which party have talked the most about the political concept of “finlandization”? **Answer:** Mr. Georg Ehrnrooth, Kansallinen Kokoomus party
3. **Question:** Who has given most often regular speeches (varsinainen puheenvuoro in Finnish) at different times? **Answer:** Mr. Veikko Vennamo, SMP Party, over 12 600 speeches in 1945–1987 in total
4. **Question:** Which government party representatives and parties do the opposition MPs refer to in their speeches? **Answer:** This can be determined and visualized by extracting and linking mentions of MPs in the speech texts, which enables, for example, network analyses of references [34].
5. **Question:** Who MP most often interrupts (with an interjection) the speeches of ministers Annika Saarikko, Krista Kiuru and Sanna Marin in the current parliament? **Answer:** Mr. Ben Zyskovicz, in the cases of Saarikko and Kiuru 46% of the interruptions are due to him, and the case of Marin 39%.

The answers to this kind of questions, for example, can be determined computationally with the help of PARLIAMENTSAMPO's data, LOD service, and portal as discussed in [5]. This system is based on the “Sampo Model” [35] that 1) explicates principles for collaborative LOD production, based on a shared ontology infrastructure, and 2) principles for user interface design where semantic faceted search and browsing is seamlessly integrated with data-analytic tools needed in DH research [36]. This approach arguably suggests for a paradigm change of DH on the Semantic Web [37].

---

<sup>8</sup>Open data services of PoF: <https://avoindata.eduskunta.fi/#/fi/home>

<sup>9</sup>FAIR Data initiative: <https://www.go-fair.org/>

## 3.2. Datasets and Knowledge Graphs

The data in PARLIAMENTSAMPO consists of two core datasets:

1. **Speeches of Plenary Sessions** This dataset contains all speeches of the Finnish parliamentary plenary debates since the PoF was established in 1907, totalling ca. 985 000 speeches by the end of 2022. This data have been transformed into a Linked Data Knowledge Graph (KG) [2] called S-KG. In addition, the speech data have been published as CVS tables and using the XML TEI-based format Parla-CLARIN<sup>10</sup>. In addition, a subcorpus in ParlaMint format was created as part of the Pan-European ParlaMint II project<sup>11</sup>.
2. **Ontology about the MPs and PoF** A knowledge graph called P-KG has been created for representing biographical data about all ca. 2800 Finnish MPs and other speakers in plenary sessions from the same time period (1907–2022), and about related parties, groups, organizations, and other entities of the PoF. [3] We will call the data model of the P-KG as the PoF Ontology.

The data transformation pipeline of PARLIAMENTSAMPO contains accordingly two branches: one for transforming the speeches [2, 38, 6] and one for creating the ontologies about the politicians and PoF [3] involved. In the next section these pipelines are explained especially from a re-use of results point of view, including:

1. Publication: CSV tables of the speeches in the minutes as data dumps
2. Publication: Parla-CLARIN formatted XML files of the speeches, including a smaller subcorpus of ParlaMint-formatted data
3. Publication: LOD version of the speeches as RDF 1.1. Turtle<sup>12</sup> files interlinked with the knowledge graph about the speakers and other entities of the PoF and their activities
4. Publication: CSV file parliamentMembers.csv describing the data about the MPs and other speakers at the PoF
5. Publication: LOD data service with a SPARQL<sup>13</sup> endpoint available on the Semantic Web

The CSV files contain the speeches extracted from the minutes as text with simple basic literal metadata, such as the name of the speaker, his/her party, session of the speech, and date. In Parla-CLARIN XML-format more detailed structured information is provided, such as data about the MPs, with internal XML identifiers for cross-referencing the data. ParlaMint is an extension of Parla-CLARIN where also linguistic information about the texts is made available, such as part-of-speech data and recognized named entities. The most versatile format for publishing speech and PoF data is linked data in RDF Turtle format where URI identifiers are used systematically within the documents and also globally, and concepts used are defined

---

<sup>10</sup>Parla-CLARIN homepage: <https://github.com/clarin-eric/parla-clarin>

<sup>11</sup><https://www.clarin.eu/parlamint>

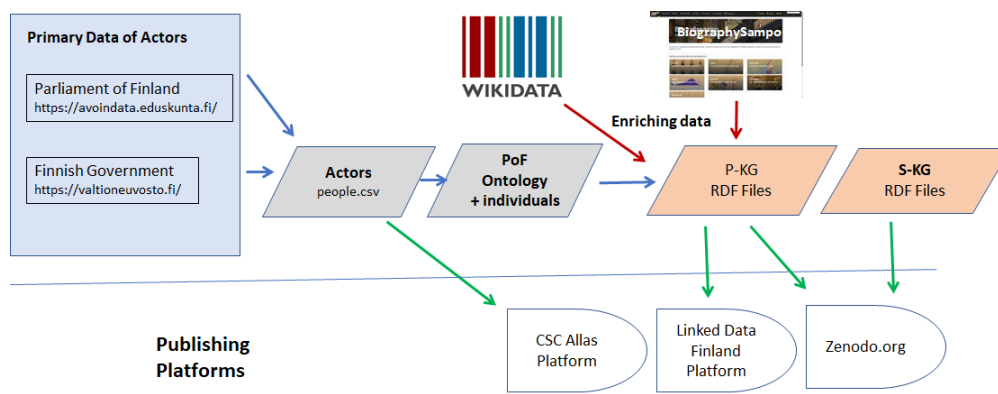
<sup>12</sup><https://www.w3.org/TR/turtle/>

<sup>13</sup>SPARQL qUery language for RDF data: <https://www.w3.org/TR/sparql11-query/>

in terms of ontological constructs of the PoF Ontology and related ontologies of the Finnish ontology infrastructure [39] and beyond. This kind of linked data can be published not only as data dumps but as live LOD services on the Web using the Linked Data publishing principles and best practices of W3C<sup>14</sup> [40, 41]. These include, e.g., support for minting and resolving URIs, browsing the KGF, and provision of a SPARQL Application Programming interface (API) for querying and re-using the data.

We next present how the prosopographical and ontological data about MPs and other entities of the PoF were extracted, enriched and published. This data is needed when transforming the speech data. Finally, the LOD version of both datasets and its data service on the Web are explained.

## 4. An Ontology of MPs and the Parliament of Finland



**Figure 1:** Pipeline of transforming data about the MPs and other speakers of the plenary sessions into the ontology of the PoF

This section describes PoF Ontology, a data model and KG of the Finnish MPs and other speakers in the plenary sessions PoF since 1907, parties, groups and organizations involved, and parliamentary events and proceedings in time and place. This ontology with its individuals constitute the P-KG that is used in the metadata descriptions and annotations of parliamentary speeches of the related speech S-KG.

### 4.1. How the Parliament of Finland Works

The organization and activities of the PoF are documented in [42]. Legislation procedures in PoF can be initiated today by a *government bill* (hallituksen esitys), by a *parliamentarian's proposal* (kansanedustajan esitys) of an MP, or as *citizen's initiative* (kansalaisaloite). The process starts with *referral discussion* (lähete keskustelu) that sends the bill to committee in whose expertise domain the bill/proposal/initiative is related to. The committees consist of

<sup>14</sup><https://www.w3.org/standards/semanticweb/data>



17 or 21 MPs and vice members. At the moment, there are 16 permanent committees in PoF. Based on a report of the committee the parliament then has a *first discussion* about the legal document in question after which still some modifications to the document can be made. Later on there is a *second discussion* where the document is finally either accepted or rejected. There are hence usually three plenary sessions where a document is discussed. The minutes about them are used in PARLIAMENTSAMPO. Our data does not contain the committee reports or other background materials. However, these have been published by PoF electronically or in print as part of the minutes, and are also processed in the Lakitutka project. PARLIAMENTSAMPO data can include links to Lakitutka contents when available. PARLIAMENTSAMPO data does not contain the final statutes (legislation) either, but they are being published by the Finlex and Lawsampo systems to whom links can be provided in PARLIAMENTSAMPO. In addition to legislative matters, the PoF discusses in plenary sessions also many other matters, such as the state budget proposals and interpellations of the opposition parties.

The work at the PoF involves people, committees, parties, and other organizations in different roles and in relation to the discussions. Furthermore, the organizational structure has evolved in time. Creating an overarching ontology over different times is a challenge due to the dynamic nature of the PoF: lots of parties, groups and other organizational units, have been established, restructured, and vanished since 1907. Reassembling the history of the PoF from the documents available was deemed infeasible, and we therefore created the ontology in a data-driven fashion based on the data available concerning the MPs and other speakers in the plenary sessions and governments making the proposals. This data included, e.g., information about the memberships of the MPs in different parties and committees, which data could be used for creating classes for the PoF ontology and for populating the ontology with instance data. In addition to the MP database available from the PoF, additional data was available from the open data sources of the Government<sup>15</sup>, BiographySampo.fi [43, 44], and Wikidata<sup>16</sup>.

The most important data source was the database of MPs provided by the PoF. It contains in structured custom XML-form basic biographical data about all MPs, such as date and place of birth, periods of time as an MP, electoral districts, memberships in parties, committees, other groups, and organizations, and publications of the MPs. From this data it was possible to extract ontological classes for PoF Ontology, such as electoral districts, parties, and committees, and at the same time populate the ontology with instances of people, committees, and other classes. This data was then extended with data about ca additional 200 speakers in the PoF that have not been MPs and therefore were not included in the MP database of the PoF. Furthermore, the speaker data was enriched with additional information available from BiographySampo.fi and Wikidata regarding, e.g., family relations, events of personal biographical history, and photographs.

## 4.2. Data Model for Parliamentary Actors, Groups, and Events

The data model of the PoF Ontology is presented in Fig. 2. It is based on the Bio CRM [45] ontology, an extension of CIDOC CRM<sup>17</sup> for representing biographical information based on

---

<sup>15</sup><https://valtioneuvosto.fi>

<sup>16</sup><https://wikidata.org>

<sup>17</sup><https://cidoc-crm.org>

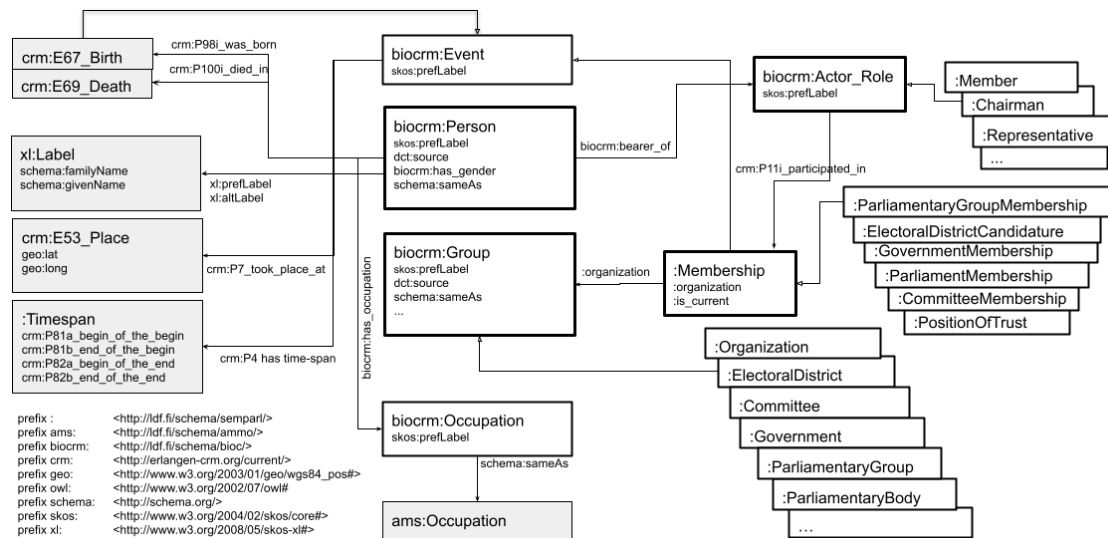


Figure 2: PoF Ontology data model [3] based on Bio CRM

role-centric modeling. Bio CRM makes a distinction between attributes, relations, and events, where entities participate in different roles in a qualified manner. The namespaces used in the model are described in the figure on the left.

The key idea of the model is to represent an actor's activities as a sequence of events (*bioc:Event*) in places (*crm:E53\_Place*) and in time (*:Timespan*) with the actors (*bioc:Person*) participating in different roles (*bioc:Actor\_Role*), such as *:Member*, *:Representative*, etc.

There are almost 200 different roles in use in the PoF Ontology. The data model has been populated by the MP database and related sources as well as by using a set of external domain ontologies, such as places based on the ontology YSO Places<sup>18</sup>, groups and organizations (harvested from the data), and vocations based on the AMMO ontology [46].

Table 1 summarizes the number of instances of the main classes of the data model of Fig. 2, and Table 2 lists the number of different event types extracted.

### 4.3. Data Quality and Validation

The ontology was created in a data-driven fashion. This means that if the data misses something, say the membership of an MP in a particular committee at a time, then the list of members in that committee instance is incomplete. It is known that the data is not fully complete. For example, the MP database for some old committees record only their chairs, not ordinary memberships. Checking and analysing possible missing data has no been done systematically afterwards; it is assumed that the database is complete in this sense and that the user is aware about the fact that this may not always be the case. Validation could be done based on historical sources that, e.g., provide lists of members in different committees in different times if such data can be found.

<sup>18</sup><https://finto.fi/yso-paikat/en/>



**Table 1**  
Resources

Resource type	Count
Timespan	10733
Label	6115
Place	4543
Person	2828
Publication	1727
Vocation	1456
School, College	670
Parliamentary Group	89
Government	76
Committee	54
Organization	54
Electoral District	46
Party	44
Parliamentary Body	38
Ministry	12
Affiliation Group	10

**Table 2**  
Events

Event type	Count
Career Event	14756
Position of Trust	12788
Committee Membership	6669
Municipal Position of Trust	4740
Event of Education	3722
Birth	2828
Parliamentary Group Membership	2280
Electoral District Candidature	2211
Death	2071
Government Membership	1637
Governmental Position of Trust	1615
Affiliation	1397
Parliament Membership	966
Honourable Mention	537
International Position of Trust	364
Membership Suspension	25

For validating the transformed data, the data model and its integrity constraints can be presented in a machine-processable format using the ShEx Shape Expressions language<sup>19</sup>. We have made initial validation experiments with the PyShEx<sup>20</sup> validator. Based on the experiments, we have identified some errors both in the schema and the data. We plan a full-scale ShEx validation phase integrated in the data conversion and publication process to spot and report errors in the dataset.

#### 4.4. PoF Ontology Available Online

The PoF Ontology with RDF data are available as RDF Turtle files on Zenodo.org using the CC BY 4.10 license:

<https://doi.org/10.5281/zenodo.7636420>

In addition to the RDF files, the central CSV data file `people.csv` about the MPs and other speakers in the plenary session are available at the Allas store at:

<https://a3s.fi/parliamentsampo/actors/csv/index.html>

Furthermore, the linked data is available in the LDF.fi platform at

<sup>19</sup><https://shex.io>

<sup>20</sup><https://github.com/hsolbrig/PyShEx>

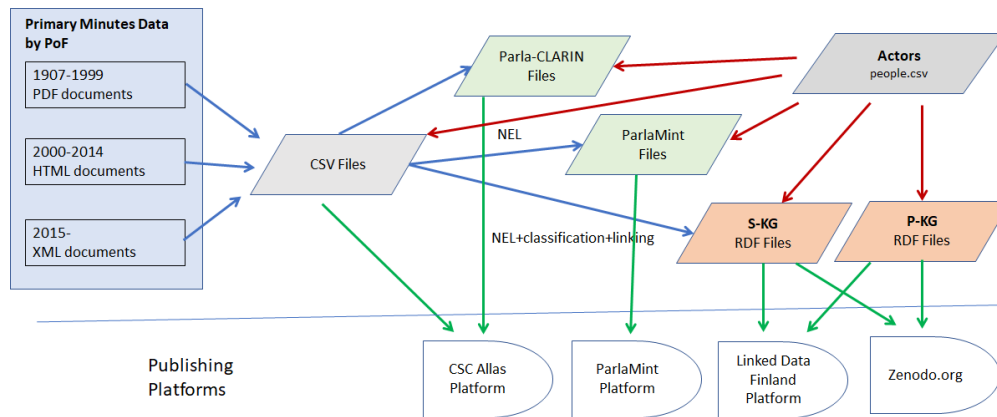
<https://www.lda.fi/dataset/semparl>

as separate graphs interlinked with the S-KG in a SPARQL endpoint.

The data can be downloaded also through the PARLIAMENTSAMPO portal that includes tools for CSV download, too. In this way the CSV data can be filtered before downloading using the faceted search of for the portal <https://parliamentisampo.fi>. For example, only people of a party during a period of time can be downloaded.

## 5. Speech Data of Plenary Sessions

Plenary discussions in PoF consist of *sessions* where particular topics or proposals, such as a bills of government, are discussed. Each session consists of a series of speeches of different types.



**Figure 3:** Pipeline for transforming the minutes of plenary sessions into speech data

Fig. 3 illustrates the process used for transforming the minutes of plenary sessions into datasets and services on different publishing platforms. The data is first transformed into simple literal data CSV tables that are published using the CSC Allas data store<sup>21</sup>. The CSV format can be of use for DH researchers developing and using their own tools, and this data also serves as the primary source for publishing more developed versions the the data. The CSV data is enriched into Parla-CLARIN XML TEI<sup>22</sup> form that includes, e.g., identifiers for speakers and into ParlaMint format where additional linguistic annotations pertaining to, e.g., named entities in the texts are available. A ParlaMint subcorpus has been published as part of the larger collection European ParlaMint corpora provided by the ParlaMint platform<sup>23</sup> [47]. The

<sup>21</sup><https://a3s.fi/parliamentsampo/speeches/csv/index.html>

<sup>22</sup><https://tei-c.org/>

<sup>23</sup><https://www.clarin.eu/parlamint>

semantically richest publication form of the data is as a semantic net in RDF form, combining the KGs of speech data and the related KG of prosopographical data and the PoF, enriched with additional data from several external sources. This data is available not only as data dumps on Allas and Zenodo.org but also as a LOD service on the Linked Data Finland platform<sup>24</sup> [48], including a SPARQL endpoint, content negotiation of URIs, linked data browsing, and other services. When enriching the CSV tables into XML and RDF formats, the interruption markup in the speeches is extracted from the text and transformed into structured forms that can be used in data analyses.

### 5.1. Speeches as CSV Tables

In the process the minutes were first transformed into simple textual CSV files. The rationale for producing and publishing CSV tables is that they can be used easily by spreadsheet programs for analysing the data or by using computational methods. From a computational point of view, they can be created automatically because no advanced data processing, such as named entity linking, is included in process. The CSV is also a useful format for checking and correcting manually errors in the results of data transformations, such as OCR errors. An example of another national parliament corpus that makes use of CSV and TSV formats is the Talk of Norway (1998–2016) [49].

The speech CSV data comes from three sources and in three different formats depending on the time of the sessions:

1. **Corpus 1907–1999** The older plenary session minutes were available only in PDF format<sup>25</sup>. These documents, often over 1000 pages long, have been created by PoF who digitized the printed minutes books of all plenary sessions. In order to extract their textual contents, we re-OCR'd the PDF documents with more accurate results than before [6]. Long documents were split into 1–8 separate PDF files, each containing the minutes for several plenary sessions. The extracted texts were structured by Python scripting into a set of CSV tables.
2. **Corpus 1999–2014** From halfway 1999 to the end of 2014, the minutes were available also in HTML format at the PoF's web pages<sup>26</sup>. The HTML documents were transformed into CSV tables.
3. **Corpus 2015–** From 2015 on the plenary session minutes are available also based on a custom-made XML schema from the *Avoim eduskunta* API<sup>27</sup>. These XML documents were transformed into CSV tables.

Each source format 1–3 differs in terms of the metadata included in the minutes. However, all formats contained the following core metadata elements about the session, speaker, and the speech: 1) Session data: session identifier, session date, session ending and starting times 2)

---

<sup>24</sup><https://ldf.fi>

<sup>25</sup>Parliament of Finland open data: <https://avoindata.eduskunta.fi/#/fi/digitoidut/download>

<sup>26</sup>Available at: <https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx>

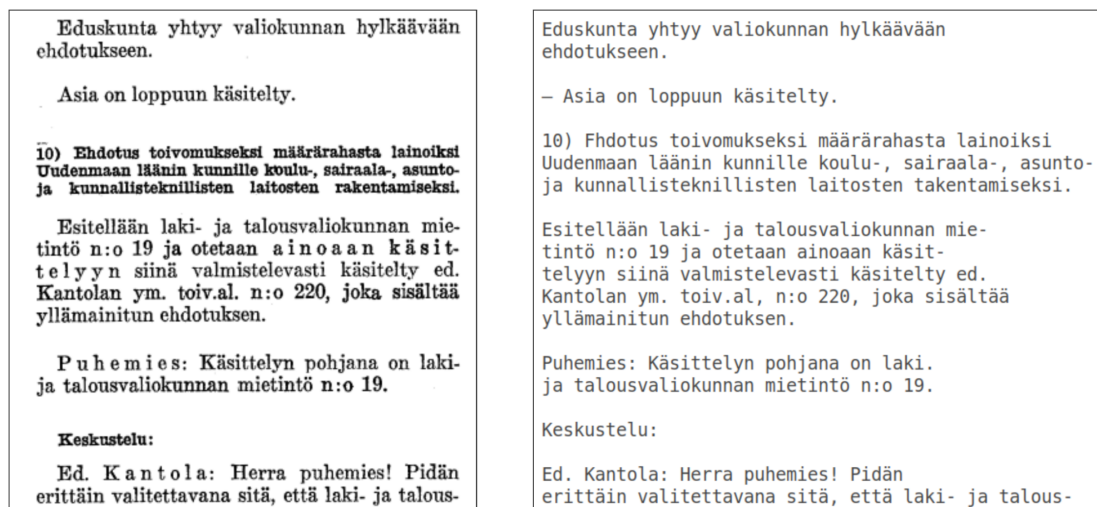
<sup>27</sup>Open PoF API: <https://avoindata.eduskunta.fi/#/fi/home>

Speaker data: last name, speaker's role/title 3) Speech data: speech content, speech type, related documents, and debate topic.

Each table row contains an individual speech with the content and metadata elements represented in columns.

### CSV Data of Years 1907–1999

Minutes of the plenary sessions in 1907–1999 were available only as images in PDF format. Fig. 4 shows an example of original minutes for a plenary session on the left. In general, the minutes consist of items (or topics), marked here in bold (except the row *Keskustelu*). The item header is followed by 1) a possible list of related documents, 2) chairman's opening comments, 3) possible debate section marked by *Keskustelu*: (*debate/conversation*) and 4) finally a decision and a closing statement. Also later minutes available in structured HTML and XML formats mostly follow this layout and logic.



**Figure 4:** OCR example. On the left a part of the original PDF-document; on right the same part with recognized text. [38]

The structure of CSV table 1907–1999 is explained in Table 3.

### CSV Data of Years 2000–2014

The structure of the CSV tables based HTML-formatted minutes in 2000–2014 is fairly similar to those in 1907–1999 discussed above.

### CSV Data of Years 2015–

Starting from 2015 the minutes are published as XML files; the corresponding CSV table format contains the following columns for metadata about speeches: party, topic, content, speech\_type,

**Table 3**

Metadata elements (columns) in annual CSV files 1907–1999. Name space ps means <http://ldf.fi/semparl/>.

Column	Meaning	Example
speech_id	Unique identifier of the speech	1925_1_2
session	Identifier of the session of the speech	1/1925, 2/1925, ...
date	Date of the session	1925-02-03
start_time	Start time of the session	10.15
end_time	End time of the session	12.15
given	Given name of the speaker	Kyösti
family	Family name of the speaker	Kallio
role	Role of the speaker	Puhemies (Speaker)
party	Party of the speaker	KESK
topic	Topic of the session and a list of related documents	1) Ehdotus ...
content	Speech text (OCR)	Kun ed. Kallio ...
speech_type	Type of the speech	Puheenvuoro, PuhemiesPuheenvuoro
mp_uri	URI identifier of the speaker	ps:people/p910662
gender	Gender of the speaker	Male, Female
birth	Day of birth of the speaker	<a href="http://ps:groups/Q506591">http://ps:groups/Q506591</a>
party_uri	URI identifier of the party	ps:groups/g8162617958878698175
parliamentary_role	Parliamentary role of the speaker	Oppositio puolue (opposition), Hallitus puolue (government)
group_uri	URI of the parliamentary group	ps:groups/Q499029
link	Link to the original PDF document	<a href="https://s3-eu-west-1.amazonaws.com/eduskunta-asiakirja-original-documents-prod/suomi/1925/PTK_1925_1.pdf">https://s3-eu-west-1.amazonaws.com/eduskunta-asiakirja-original-documents-prod/suomi/1925/PTK_1925_1.pdf</a>
lang	Language of the speech	fi, sv, fi:sv
name_in_source	How the speaker is addressed textually	Ed. Tanner, Ed. Kallio, Puhemies
page	Starting page of the speech in the PDF document	13

status, version, link, lang, name\_in\_source, speaker\_id, speech\_start, speech\_end, speech\_status, and speech\_version. See the Allas repository for more information about the metadata elements.

### Markup in Text Content

In addition to metadata about a speech, the speech text itself contains mark-up metadata about possible interruptions of the speech using special bracketed notation. The interruptions are made by other people during the speech and in many cases the minutes also tell who made the interruption. For example: “... nostamiseksi [Arto Satosen välihuuto] hallitusohjelman ... ” (*Arto Satonen’s interruption* in English). In the CSV data the marked interruptions are left intact in texts. However, during the next data processing steps they were extracted as new metadata that can be used in data analyses. In data 1907–1999 interruptions are marked with parantheses “(*interruption text*)” and after that with brackets “[*interruption text*]”.

The practises on how minutes of plenary sessions should be recorded are described in a lengthy 147-page document of the Minutes Office of PoF (“pöytäkirjatoimisto” in Finnish) [50]. It is not fully known what kind of changes in practice there have been at different times. These changes may have implications of data analyses in some cases. For example, in 2021 it was decided that if the Speaker (“puhemies” in Finnish) only gives the floor to the next speaker without other content in his/her speech, then this is not recorded as a distinct speech of the Speaker for simplicity. If the number of all kind of speeches of different kinds in different times is analyzed, this change in recording practise of course skews results statistically.

## Automatic Updates of CSV tables

The CSV data of the past years is stable but can be updated on an irregular basis when, e.g., OCR errors etc. are found in the data. Information about the updates will be stored in the `readme.txt` file stored in the same folder as the CSV files.

As new minutes are published by the PoF on their data service, the CSV table of the current year is updated automatically on a daily basis with the new speeches.

## CSV Tables Available on the Web

The CSV tables are published as files that were created on parliamentary session basis, one file per parliamentary session (valtiopäivät) with the name `speeches_YEAR[_N].csv`, where `YEAR = 1907, 1908, ... and [_N]`, `N = II | XX` is optional. For example, speeches from 1925 are in the file `speeches_1925.csv`. However, occasionally there have been two parliamentary sessions referring to the same calendar year<sup>28</sup>. For example the speeches from the first parliamentary session of 1918 are in the file `speeches_1918.csv` and speeches from the second parliamentary session are in `speeches_1918_II.csv`. The years 1915 and 1916 are missing because the PoF did not convene then due to the World War I. In 1917 between first and second parliament, two unofficial meetings were held. These meetings have been given (originally lacking) order numbers for the sake of itemization. Files containing data from these meeting are marked by `_XX`. The CSV tables are available openly with the CC BY 4.0 license at the Allas data repository of CSC Ltd:

<https://a3s.fi/parliamentsampo/speeches/csv/index.html>

This folder includes 1) a zip file that contains the CSV data files of all parliamentary sessions, 2) the parliamentary session files as separate CSV files, and 3) a link to documentation. The last file of the current parliamentary session is updated daily.

## 5.2. Speeches in Parla-CLARIN and ParlaMint Formats

The XML TEI-based Parla-CLARIN [47] schema is an attempt to define a common XML-based annotation model for parliamentary debates on an international level.<sup>29</sup> For example, the Slovene parliamentary corpus siParl (1990–2018) has been encoded with the Parla-CLARIN schema [14]. Currently, the Parla-CLARIN schema is implemented in the Clarin ParlaMint project<sup>30</sup>, which establishes a comparable and interoperable corpus of almost twenty national parliamentary corpora for comparative research. This format is a specialization of Parla-CLARIN extending it with, for example, linguistic annotations and named entity mentions.

Parla-CLARIN format includes not only speeches but also means for representing data about the context of the debates including data about the speakers, parties, related organizations, and

<sup>28</sup>Due to the Government resigning prematurely and thus starting a new parliamentary session

<sup>29</sup>See: <https://www.clarin.eu/blog/clarin-parlaformat-workshop>

<sup>30</sup><https://github.com/clarin-eric/ParlaMint>



places in systematic way using XML identifiers for cross-reference. A benefit of using XML-based formats is the possibility of validating documents syntactically based on their schema definition.

The Parla-CLARIN version of the PARLIAMENTSAMPO speeches is available at the Allas data store at:

<https://a3s.fi/parliamentsampo/speeches/xml/index.html>

The ParlaMint subcorpus is under validation and will appear later in the ParlaMint data repository<sup>31</sup>.

### Publication as Linked Open Data

The LOD version of the speech data was created from the CSV tables, too [2, 38]. The latest corpus 2015– has been annotated semantically using Natural Language Processing (NLP) techniques as discussed in [51]:

1. **Named Entity Linking.** Mentions of the MPs and places were extracted, disambiguated semantically, and linked to corresponding resources with URIs in the PoF Ontology data. These annotations facilitate, e.g., network analyses on MPs and parties based on mutual references in speeches as discussed in [34].
2. **Automatic keyword annotation.** Finnish NLP technology was applied also for annotating the speeches automatically using the YSO ontology<sup>32</sup> [52] of the National Library of Finland and the Annif automatic annotation tool<sup>33</sup> [53]. Ontology-based keywords facilitate semantic search and content-based analyses of the speeches. The data includes also keywords extracted using the TF-IDF method.
3. **Automatic library classification.** The EKS subject headings<sup>34</sup> vocabulary of the Library of PoF was transformed into a SKOS<sup>35</sup> ontology, and the sessions were indexed automatically based this. EKS subject headings annotations facilitate hierarchical topical classification of the sessions and their speeches.
4. **Linguistic data.** The data also includes additional linguistic analysis data, such as lemmatized versions of the speech texts.

The NLP-based annotations have been published as part of the PARLIAMENTSAMPO RDF Turtle data dump in Zenodo.org<sup>36</sup> and as linked open data on the Linked Data Finland platform<sup>37</sup>.

<sup>31</sup>See the current ParlaMint 2.1 version: <http://hdl.handle.net/11356/1432>

<sup>32</sup><https://finto.fi/yso/fi/>

<sup>33</sup><https://annif.org/>

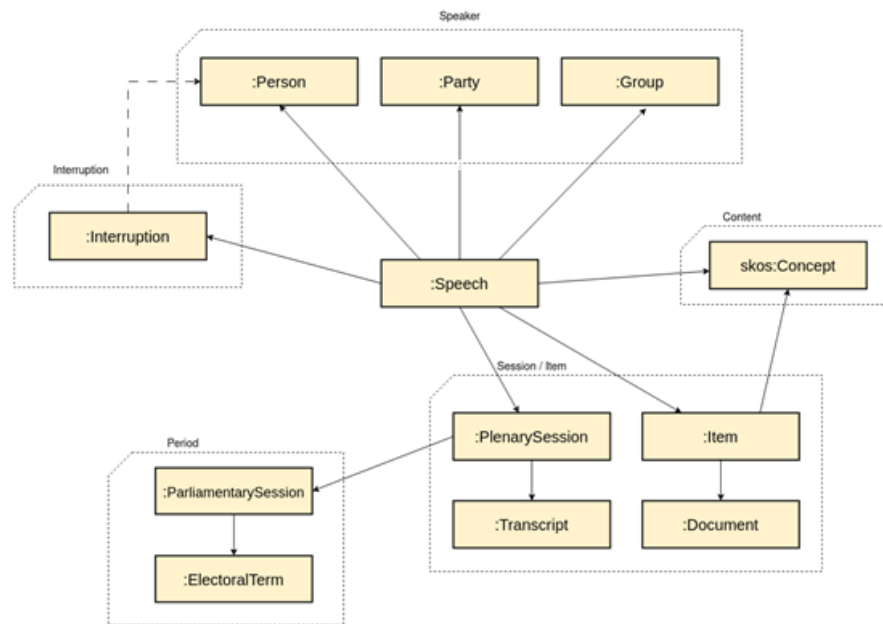
<sup>34</sup><https://www.eduskunta.fi/kirjasto/EKS/>

<sup>35</sup>Simple Knowledge Organization System: <https://www.w3.org/TR/skos-reference/>

<sup>36</sup><https://doi.org/10.5281/zenodo.7636420>

<sup>37</sup><https://www.ldf.fi/dataset/sem parl>

## Data Models for Speeches and Their Annotations



**Figure 5:** Data model for speech data in the default namespace <https://ldf.fi/schema/sempar/>

The data model of speech data is depicted in Fig. 5; additional documentation can be found in [2]. The speeches of the latest and best quality dataset 2015– have been annotated with extracted named entities, keywords, and EKS categories, and the data also includes lemmatized versions of the speeches. The datamodel for these annotations can be seen in Fig. 6. More documentation about these data models can be found using the namespace URL in a browser.

## 6. Using the PARLIAMENTSAMPO Data

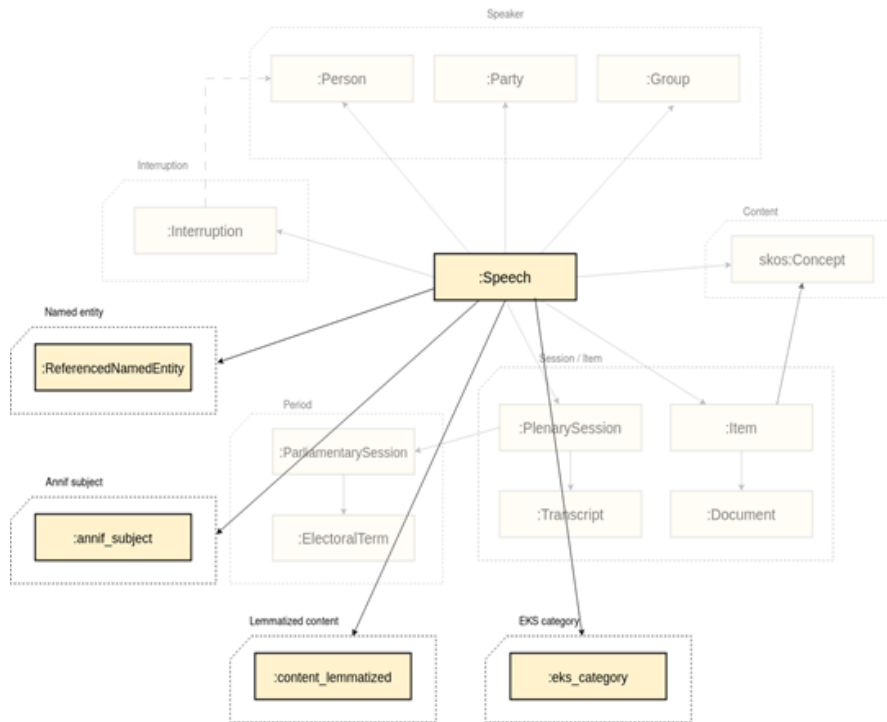
This section discusses briefly different ways of using the PARLIAMENTSAMPO infrastructure in research.

### 6.1. Exporting the Data for External Use

A simple way for a researcher to use PARLIAMENTSAMPO data is to download data from the data services for local use and then apply one's favourite tools for data analysis, such as spreadsheets, R<sup>38</sup> environment for statistical analysis, or Gephi<sup>39</sup> for network analysis. For filtering out subsets of interest in the big data, SPARQL querying can be used in flexible ways. It is also

<sup>38</sup><https://www.r-project.org>

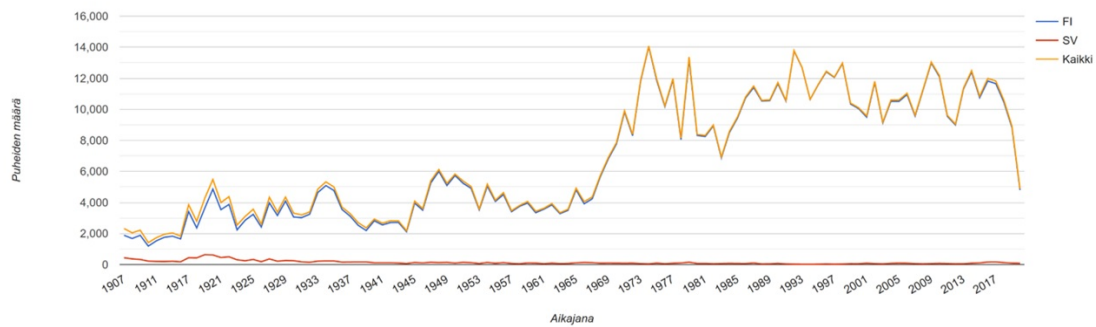
<sup>39</sup><https://gephi.org>



**Figure 6:** Data model for the linguistic annotations of speech data 2015– in the default namespace <https://ldf.fi/schema/semparl/>

possible to install a local SPARQL server environment for linked data on one’s own computer, for example Fuseki<sup>40</sup>, which is also used in the LDF.fi service. The materials in the LDF.fi service are published using container technology (i.e., Docker<sup>41</sup>), which means that installing the data, the server, and possible versioned software packages is automatic and effortless.

An example of using the PARLIAMENTSAMPO data externally is reported in [25]. 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 the EKS subject headings thesaurus, 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.



**Figure 7:** Number of speeches in different languages (y-axis) on the timeline (x-axis).

## 6.2. Querying the Endpoint and Studying Results

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. 7 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 editor<sup>42</sup> [54], which can be used to edit SPARQL queries, target them to an online SPARQL endpoint, and to show the results using pre-implemented visualizations.

## 6.3. Data-analysis by Scripting

The PoF data can be examined computationally, for example, using Python scripting and Jupyter notebooks in the Google Colab<sup>43</sup> 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. An example anaöysis of using GÖoogle Colab is presented in Table 8.

## 6.4. Using the PARLIAMENTSAMPO Portal

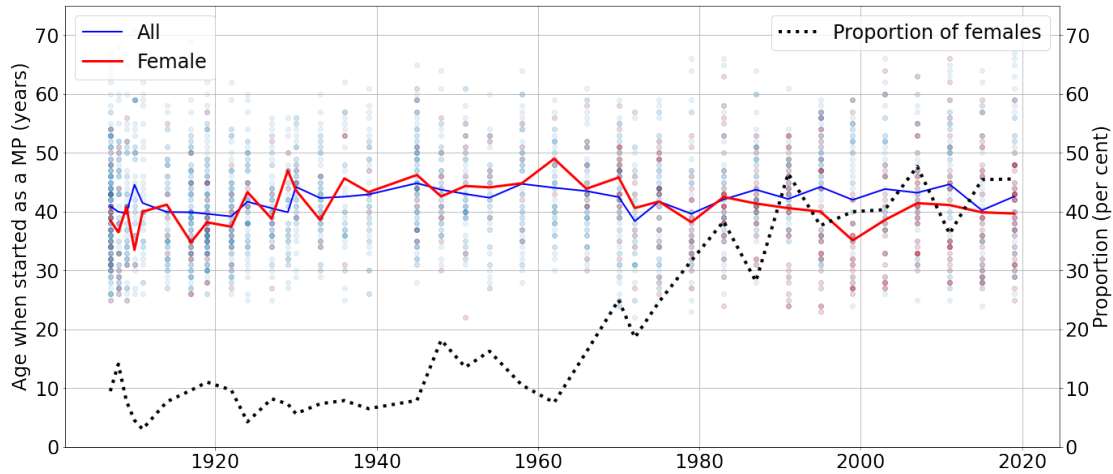
The PARLIAMENTSAMPO portal, based on the Sampo model [35] and the Sampo-UI framework [36], demonstrates how the SPARQL data service can be used for developing applications for DH research. In the portal, the data can be filtered using faceted search [55] based on ontologies, and the results can then be analyzed with the help of seamlessly integrated visualization and

<sup>40</sup><https://jena.apache.org/documentation/fuseki2/>

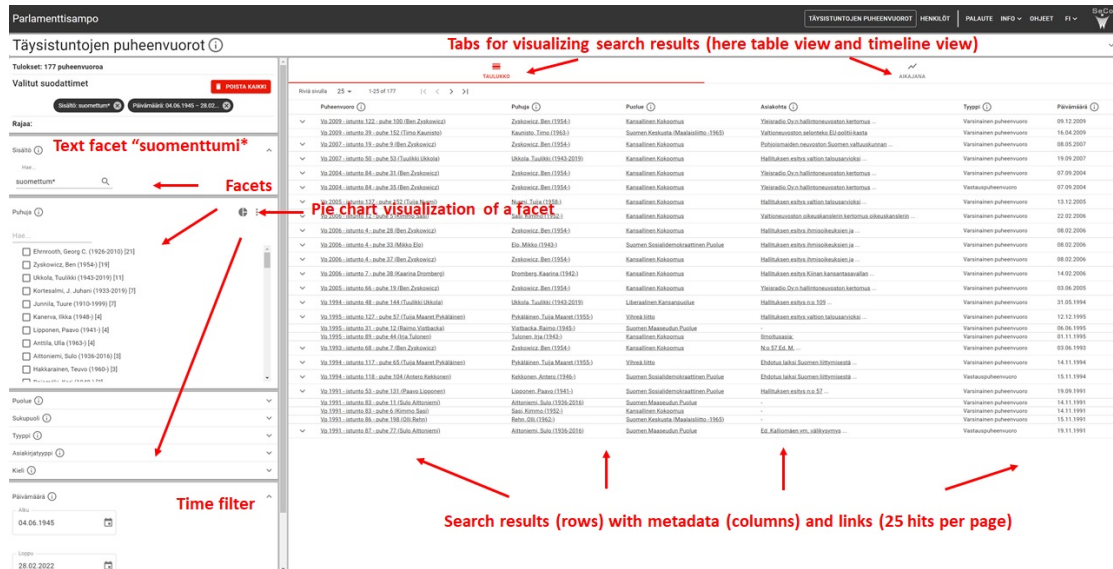
<sup>41</sup><https://www.docker.com>

<sup>42</sup><https://yasgui.triply.cc>

<sup>43</sup><https://colab.research.google.com>



**Figure 8:** Annual starting age of new MPs, male and female, and relative proportion of women in the PoF on a timeline



**Figure 9:** Using faceted search to filter out speeches of interest.

data analysis tools. The data can be accessed along application perspectives for studying 1) speeches of different times and 2) MPs and other speakers. For example, in Fig. 9, 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 English) 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. A number of pre-implemented data analysis tools and visualizations, similar to those shown in the figures above, have been integrated into the application perspectives of the PARLIAMENTSAMPO portal<sup>44</sup>.

## 7. Discussion

The datasets of PARLIAMENTSAMPO presented in this paper make it possible to utilize the speeches of the plenary debates of PoF as well as data about the speakers and other entities in the PoF in DH research. For the first time, a machine-readable data corpus covering the whole history of the PoF since 1907 including nearly million speeches and over 2800 parliamentarians has been created and published openly as data and data services services. Usefulness of the datasets and services has been demonstrated by using them in data analyses and by implementing the PARLIAMENTSAMPO portal in use that demonstrates how the data can be used for application development and data analyses.

In traditional close reading, the researcher is forced to delimit the data studied on, e.g., temporal or thematic grounds. Digital methods applied to big data, such as PARLIAMENTSAMPO, make it possible to study political culture and language without such limitations. For example, new themes and topics can be identified automatically or semi-automatically (e.g., [56, 57]) and the language of politics and its long-term changes can be studied (e.g., [58, 59, 60, 61, 62, 63, 31]). Furthermore, by linking the data to data about the parliamentarians and their activities and other entities in the PoF and beyond, the social contexts of language users, such as education, gender, age, and social networks can be studied (e.g., [64, 34]).

**Acknowledgements** Thanks to Esko Ikkala, Mikko Koho, and Minna Tamper for their contributions in the ParliamentSampo project earlier. Fruitful collaborations and discussions with Kimmo Elo, Jenni Karimäki, and Anna Ristilä of the University of Turku, Center for Parliamentary Studies, regarding the use cases and research of parliamentary data are acknowledged. PARLIAMENTSAMPO is based on the open data from the PoF: thanks to Ari Apilo, Sari Wilenius, and Päivikki Karhula of POF for collaborations. Our work was funded by the Academy of Finland in the projects Semantic Parliament<sup>45</sup> and FIN-CLARIAH<sup>46</sup>, by CLARIN.eu in the ParlaMint II project<sup>47</sup>, and is also related to the EU project InTaVia<sup>48</sup> and the EU COST action Nexus Linguarum<sup>49</sup> on linguistic linked data data resources and analysis. The project uses the computing resources of the CSC – IT Center for Science.

---

<sup>44</sup>Available at: <https://parlamenttisampo.fi>

<sup>45</sup><https://seco.cs.aalto.fi/projects/semparl/>

<sup>46</sup><https://seco.cs.aalto.fi/projects/fin-clariah/>

<sup>47</sup><https://www.clarin.eu/parlamint>

<sup>48</sup><https://intavia.eu>

<sup>49</sup><https://nexuslinguarum.eu>



## References

- [1] C. Benoît, O. Rozenberg (Eds.), *Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures*, Edward Elgar Publishing, 2020. doi:10.4337/9781789906516.
- [2] 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 fur Informatik GmbH, Dagstuhl Publishing, 2021, pp. 1–17. URL: <https://drops.dagstuhl.de/opus/volltexte/2021/14544/pdf/OASlcs-LDK-2021-8.pdf>.
- [3] 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.
- [4] 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>.
- [5] E. Hyvönen, L. Sinikallio, P. Leskinen, M. L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, Finnish parliament on the semantic web: Using parlamentsampo data service and semantic portal for studying political culture and language, in: Digital Parliamentary data in Action (DiPaDA 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, long paper, CEUR Workshop Proceedings, Vol. 3133, 2022, pp. 69–85. URL: <http://ceur-ws.org/Vol-3133/paper05.pdf>.
- [6] S. Drobac, L. Sinikallio, E. Hyvönen, An OCR pipeline for transforming parliamentary debates into linked data: Case ParliamentSampo – Parliament of Finland on the semantic web, 2022. URL: <https://seco.cs.aalto.fi/publications/2022/drobac-et-al-ocr-2022.pdf>, paper under peer review.
- [7] M. La Mela, F. Norén, E. Hyvönen (Eds.), *Digital Parliamentary Data in Action (DiPaDA 2022): Introduction*, volume 3133, CEUR WS, 2022. URL: <http://ceur-ws.org/Vol-3133/paper00.pdf>.
- [8] D. Fišer, M. Eskevich, J. Lenardič, F. de Jong (Eds.), *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, 2022. URL: <https://aclanthology.org/2022.parlaclarin-1.0>.
- [9] 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.
- [10] 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.
- [11] 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.

- [12] M. La Mela, F. Norén, E. Hyvönen (Eds.), From Early Modern Deliberation to the Semantic Web: Annotating Communications in the Records of the Imperial Diet of 1576, volume 3133, CEUR WS, 2022. URL: <http://ceur-ws.org/Vol-3133/paper06.pdf>.
- [13] D. Fišer, M. Eskevich, J. Lenardič, F. de Jong (Eds.), ParlaMint II: The Show Must Go On, ????
- [14] 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>.
- [15] 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.
- [16] 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>.
- [17] 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.
- [18] 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.
- [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] 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.
- [21] 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.
- [22] 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.
- [23] 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.
- [24] 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.
- [25] K. Elo, J. Karimäki, Luonnonsuojelusta ilmastopoliittikkaan: Ympäristöpoliittisen käsit-

- teistön muutos parlamenttipuheessa 1960–2020, *Politiikka* 63 (2021). URL: <https://journal.fi/politiikka/article/view/109690>. doi:10.37452/politiikka.109690.
- [26] 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.
- [27] A. T. J. Barron, J. Huang, R. L. Spang, S. DeDeo, Individuals, institutions, and innovation in the debates of the french revolution, *Proceedings of the National Academy of Sciences* 115 (2018) 4607–4612. URL: <https://www.pnas.org/doi/abs/10.1073/pnas.1717729115>. doi:10.1073/pnas.1717729115.
- [28] 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.
- [29] 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.
- [30] S. Simola, A century of partisanship in Finnish political speech, 2020. URL: <https://sites.google.com/site/sallasimolaecon/home/research>.
- [31] 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>.
- [32] 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>.
- [33] E. Hyvönen, Parlamenttisampo avaa eduskunnan miljoona puhetta ja kansanedustajien verkostot kaikkien tutkittaviksi, *Tieteessä tapahtuu* 41 (2023). URL: <https://seco.cs.aalto.fi/publications/2023/hyvonen-parlamenttisampo-tt-2023.pdf>.
- [34] H. Poikkimäki, P. Leskinen, M. Tamper, E. Hyvönen, Analyses of networks of politicians based on linked data: Case ParliamentSampo – parliament of Finland on the Semantic Web, in: *Semantic Web and Ontology Design for Cultural Heritage (SWODCH 2022)*, Turin, Italy, Proceedings, CEUR WS Proceedings, 2022. Accepted.
- [35] 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>.
- [36] 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.
- [37] 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.
- [38] L. Sinikallio, Eduskunnan täysistuntojen pöytäkirjojen muuntaminen semanttiseksi dataksi ja julkaiseminen verkkopalveluna (Transformation of the Debates of the Parliament of Finland into Semantic Data and a Data Service), Master’s thesis, University of Helsinki, Department of Computer Science, 2022. URL: <http://urn.fi/URN:NBN:fi:hulib-202204201707>, mSc Thesis.
- [39] E. Hyvönen, How to create a national cross-domain ontology and linked data infrastructure

and use it on the semantic web (2022). URL: <http://www.semantic-web-journal.net/content/how-create-and-use-national-cross-domain-ontology-and-data-infrastructure-semantic-web>, under review.

- [40] 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/>.
- [41] E. Hyvönen, *Publishing and using cultural heritage linked data on the semantic web*, Morgan & Claypool, Palo Alto, CA, 2012.
- [42] M. Hidén, H. Honka-Hallila, *Miten eduskunta toimii* (How Parliament of Finland works), Edita Publishing, Helsinki, 2006.
- [43] 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.
- [44] M. Tamper, P. Leskinen, E. Hyvönen, R. Valjus, K. Keravuori, *Analyzing biography collection historiographically as linked data: Case national biography of finland*, *Semantic Web – Interoperability, Usability, Applicability* (2021). URL: <https://seco.cs.aalto.fi/publications/2021/tamper-et-al-bs-2021.pdf>, accepted.
- [45] 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>.
- [46] 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/>.
- [47] T. Erjavec, M. Ogrodniczuk, P. Osenova, et al., *The ParlaMint corpora of parliamentary proceedings*, *Lang Resources & Evaluation* (2022). URL: <https://doi.org/10.1007/s10579-021-09574-0>.
- [48] 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](https://doi.org/10.1007/978-3-319-11955-7_24).
- [49] E. Laponi, M. G. Søyland, E. Velldal, S. Oepen, *The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016*, *Language Resources and Evaluation* 52 (2018) 873–893. URL: <https://doi.org/10.1007/s10579-018-9411-5>. doi:10.1007/s10579-018-9411-5.
- [50] *Kirjo – kirjaamisohjeet*, Eduskunnan kanslia, Helsinki, Finland, 2021. Guidelines for recording minutes of plenary sessions at Parliament of Finland.
- [51] M. Tamper, R. Leal, L. Sinikallio, P. Leskinen, J. Tuominen, E. Hyvönen, *Extracting knowledge from parliamentary debates for studying political culture and language*, in: S. Tiwari, N. Mihindukulasooriya, F. Osborne, D. Kontokostas, J. D'Souza, M. Kejriwal (Eds.), *Proceedings of the 1st International Workshop on Knowledge Graph Generation From Text and the 1st International Workshop on Modular Knowledge co-located with*

- 19th Extended Semantic Conference (ESWC 2022), volume 3184, CEUR WS, 2022, pp. 70–79. URL: [http://ceur-ws.org/Vol-3184/TEXT2KG\\_Paper\\_5.pdf](http://ceur-ws.org/Vol-3184/TEXT2KG_Paper_5.pdf), international Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022).
- [52] K. Seppälä, E. Hyvönen, *Asiasanaston muuttaminen ontologiaksi. yleinen suomalainen ontologia esimerkkinä finnto-hankkeen mallista (changing a keyword thesaurus into an ontology. general finnish ontology as an example of the finnto model)*, 2014. URL: <https://www.doria.fi/handle/10024/96825>.
- [53] O. Suominen, *Annif: DIY automated subject indexing using multiple algorithms*, *LIBER Quarterly* 29 (2019) 1–25. doi:10.18352/lq.10285.
- [54] L. Rietveld, R. Hoekstra, *The YASGUI family of SPARQL clients*, *Semantic Web – Interoperability, Usability, Applicability* 8 (2017) 373–383. doi:10.3233/SW-150197.
- [55] Y. Tzitzikas, N. Manolis, P. Papadakos, *Faceted exploration of RDF/S datasets: a survey*, *Journal of Intelligent Information Systems* 48 (2017) 329–364.
- [56] D. Mimno, *Topic Regression*, Ph.D. thesis, University of Massachusetts Amherst, 2012. URL: [https://scholarworks.umass.edu/open\\_access\\_dissertations/520](https://scholarworks.umass.edu/open_access_dissertations/520).
- [57] 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.
- [58] 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.
- [59] 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.
- [60] S. Purhonen, A. Toikka, “Big Datan” haaste ja uudet laskennalliset tekstiaineistojen analyysimenetelmät: esimerkkitapauksena aiheanalyysi tasavallan presidenttien uudenpuheista 1935–2015, *Sociologia* 53 (2016) 6–27. URL: <http://elektra.helsinki.fi/se/s/0038-1640/53/1/bigdatan.pdf>.
- [61] S.-M. Laaksonen, M. Nelimarkka, *Omat ja muiden aiheet: Laskennallinen analyysi vaalijulkisuuden teemoista ja aiheomistajuudesta*, *Politiikka* 60 (2018) 132–147.
- [62] 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.
- [63] J. B. Mountford, *Topic modeling the red pill*, *Social Sciences* 7 (2018). doi:10.3390/socsci7030042.
- [64] 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.