# Plenary Speeches of the Parliament of Finland as Linked Open Data and Data Services

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

[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]*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) since 1907 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 an ontology and 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 was published on February 14th 2023 on the Web using the open CC BY 4.0 license, and quickly gathered thousands of users.

**Keywords**

parliamentary studies, semantic portals, linked data, digital humanities

## 1. Parliamentary Speeches as FAIR data for Problem Solving

Openness and transparency of parliamentary work is a foundation of democracy: it is important for the voters, media, researchers of parliamentary studies and language, and the parliamentarians themselves. Minutes of plenary sessions in parliaments in particular provide lots of information about the democratic system in use, political life, language, and culture [1, 2]. This paper presents the results of transforming the plenary session minutes of the Parliament of Finland (PoF) into openly available data services, including a Linked Open Data (LOD) knowledge graph (KG) in a SPARQL endpoint, as part of the larger ParliamentSampo infrastructure [3, 4]. The data has been used in parliamentary research studies and for creating the in-use semantic portal *ParliamentSampo – Parliament of Finland of the Semantic Web*[1].

*TEXT2KG 2023: Second International Workshop on Knowledge Graph Generation from Text, May 28 – Jun 1, 2023, co-located with Extended Semantic Web Conference (ESWC), Hersonissos, Greece*
eero.hyvonen@aalto.fi (E. Hyvönen)
https://seco.cs.aalto.fi/u/eahyvone/ (E. Hyvönen)


[1]Available at: https://parlamenttisampo.fi

The minutes of the plenary sessions of PoF have been available as printed books at the Library of Parliament and Archive of Parliament, and later also through the PoF's open data service as scanned PDF documents, HTML pages or as XML documents, depending on which parliamentary sessions are in question[2]. However, they have not been published as data in accordance with modern FAIR principles in a Findable, Accessible, Interoperable and Re-usable form for searching, browsing, and data analytic applications[3]. 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, 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) and when? **Answer**: Mr. Veikko Vennamo, SMP Party, over 12 600 speeches in 1945–1987 in total

4. **Question**: Which 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 in the case of Marin 39%.

The answers to this kind of questions, for example, can be determined computationally with the help of the ParliamentSampo's data, LOD service, and portal as discussed in [5]. This system is based on the "Sampo Model" [6] 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 [7]. This approach arguably suggests for a paradigm change of Digital Humanities (DH) on the Semantic Web [8].

This paper presents the data publishing infrastructure ParliamentSampo about the speeches and politicians of the Parliament of Finland (PoF), starting from 1907 when the PoF was established. The focus is on the data about the speeches given during the plenary sessions of the PoF. To cater different user needs, this 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 the semantic portal ParliamentSampo in use on top of the LOD service SPARQL endpoint. This paper extends earlier papers about ParliamentSampo [9, 10, 11, 5, 12] by focusing on the data resources and services available on the Web that

[2]Open data services of the PoF: https://avoindata.eduskunta.fi/#/fi/home
[3]FAIR Data initiative: https://www.go-fair.org/

constitute the new, openly available infrastructure published on February 14th, 2023[4]. Within ten days after the publication, the portal had been used by ca. 3000 users [13].

In the following, related research on parliamentary speech data is first reviewed (Section 2). After this the data production pipeline of the speech data and its different outputs are explained (Section 3). Examples of using the ParliamentSampo speech data in different ways are given to illustrate the usability of the infrastructure in research (Section 4). In conclusion, results of our work are summarized (Section 5) and directions of further development discussed.

## 2. Related Work on Parliamentary Speech Data

In recent years, parliamentary debate corpora and digital parliamentary datasets have been created from the documents of both historical and contemporary parliaments [14, 15]. 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 [14, 15]. 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.[5]

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 project 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 [17]. Today, the Open Data Portal of the European Parliament provides lots of datasets as LOD and in CSV format[6]. 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 [18], the Italian Parliament data[7], and the historical Imperial Diet of Regensburg of 1576 project [19]. A key initiative for harmonization and annotation of national parliamentary corpora is the ParlaMint project part of the CLARIN infrastructure.[8] The ParlaMint project applies the TEI-based Parla-CLARIN scheme[9], and aims to create uniformly annotated multilingual parliamentary corpora with its partners. The current ParlaMint II involves 27 national parliamentary corpora [20] (see also [21]).

The minutes of the Parliament of Finland 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 metadata [9, 22]. Finnish parliamentary debates have been published as language corpora, for example by the FIN-CLARIN's

[4]Publication event homepage: https://seco.cs.aalto.fi/events/2023/2023-02-14-parlamenttisampo/
[5]See, e.g., the Lipad project and the Canadian Hansard, https://lipad.ca [16]
[6]https://data.europarl.europa.eu/en/datasets
[7]http://data.camera.it
[8]https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora
[9]https://github.com/clarin-eric/parla-clarin

Language Bank[10] [23], where the Parliamentary corpus 2008–2016 contains linguistically annotated plenary debates and also links to the session videos [24]. 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 [25]. The speeches of the Finnish parliamentarians from 1991 to 2015 have been included also in the International Harvard ParlSpeech Corpus [26], 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 [14, 15]). Metadata and annotations make 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 [27]. Parliamentary debates have been used in thematic or conceptual analyses (e.g., [28, 29, 27, 30, 31, 32]) and to study the language and the opinions of the parties or MPs (e.g., [33, 34, 35, 36, 25]). Parliamentary debates have been used in translation studies using, for example, the EuroParl Corpus[11] of the European Parliament debates.

The debates of the PoF have been employed previously in several social scientific and linguistic studies. La Mela [22], also Kettunen and La Mela [31], 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 [31]. Andrushschenko et al. [25] have used their grammatically structured corpus for selected digital humanities research cases. Simola [37] has explored the differences in political speech between parties in the long term (1907–2018), and Makkonen and Loukasmäki [38] 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. [39] 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.

## 3. Speech Data of Plenary Sessions

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. These data have been transformed into a Linked Data Knowledge Graph (KG) [9] called S-KG. In addition, the speech data have been published as CVS

[10] http://korp.csc.fi
[11] https://www.statmt.org/europarl/

tables and using the XML TEI-based format Parla-CLARIN[12]. In addition, a subcorpus in ParlaMint format was created as part of the Pan-European ParlaMint II project[13].

2. **Ontology and data 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. [10] We will call the data model of the P-KG as the *PoF Ontology*.

### 3.1. Transformation Pineline for Speech Data

The data transformation pipeline of ParliamentSampo contains accordingly two branches: one for transforming the speeches [9, 40, 12] and one for creating the ontology and data about the politicians and PoF [10] involved. In the following, the pipeline for transforming speeches from the mostly textual minutes of the plenary sessions is presented.

Plenary discussions in PoF consist of *sessions* where particular topics or proposals, such as bills of government, are discussed. Each session consists of a series of speeches of six different types (e.g., speech of the Speaker, group speech, and regular speech).

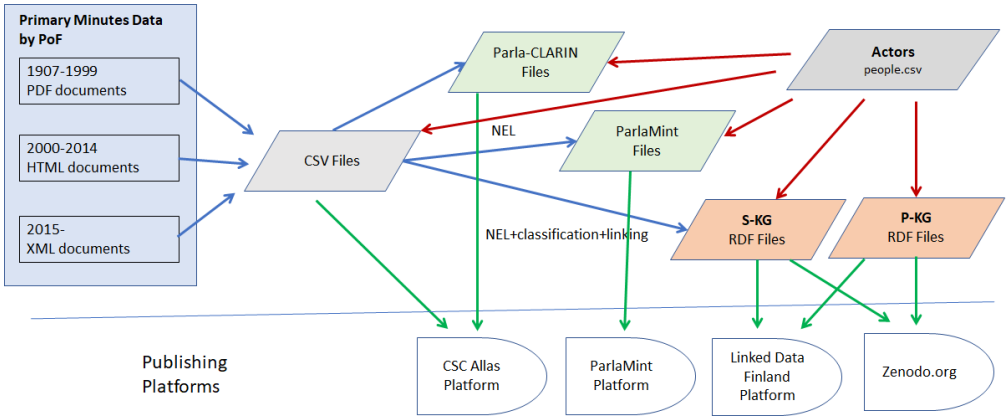



**Figure 1:** Pipeline for transforming the minutes of plenary sessions into speech data

Figure 1 illustrates the process used for transforming the minutes of the 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 national CSC Allas data store[14]. The CSV format can be of use for DH researchers developing and using their own tools, and this data publication also serves as the primary source for publishing semantically richer versions of the data. The CSV data is then enriched into Parla-CLARIN XML TEI[15] form that includes, e.g.,

[12]Parla-CLARIN homepage: https://github.com/clarin-eric/parla-clarin
[13]https://www.clarin.eu/parlamint
[14]Allas Store: https://a3s.fi/parliamentsampo/speeches/csv/index.html
[15]https://tei-c.org/

identifiers for the speakers, and into ParlaMint format where additional linguistic annotations pertaining to, e.g., named entities in the texts are explicated. Also a ParlaMint subcorpus has been created and will be published as part of the larger collection European ParlaMint corpora provided by the ParlaMint platfrom[16] [41] after it is accepted by a data validation process. The semantically richest publication form of the data is the RDF 1.1. Turtle[17] version. This publication combines the KGs of speech data and the related KG of prosopograhical data and the PoF, based on the Pof Ontology, and enriched with additional data from several external sources. This data has been published as data dumps on the Allas Store and Zenodo.org, and also as a LOD service on the Linked Data Finland platform[18] [42], 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.

### 3.2. Speeches as CSV Tables

In the transformation process the minutes are 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 and by using various 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 only exception are the URI identifiers for the speakers and parties extracted from the Actors file `people.csv` (cf. Figure 1 on the right). The CSV is also a useful format for checking and correcting 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) [43].

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

1. **Corpus 1907–1999** The older plenary session minutes were available only in PDF format[19]. These documents, often over thousand pages long, have been created by the PoF who has digitized the printed minutes books of all plenary sessions. In order to extract their textual contents, we re-OCRed the PDF documents using multilingual Deep Neural models, as presented in [12].

   Figure 2 shows the percentage of recognized words across the whole documents with the Language Analysis Command-Line Tool (LAS) [44] using the original PoF documents and our new OCR results. The new OCR results are consistently better than the original PoF version, with the biggest improvement for the material from 1920s, which is the most challenging due to poor paper quality. The words are recognized on multilingual datasets using only Finnish morphology so they do not show the absolute word accuracy rate, which is estimated to be in the 98-99 % range for Finnish text [12].

[16] https://www.clarin.eu/parlamint
[17] https://www.w3.org/TR/turtle/
[18] https://ldf.fi
[19] Parliament of Finland open data: https://avoindata.eduskunta.fi/#/fi/digitoidut/download

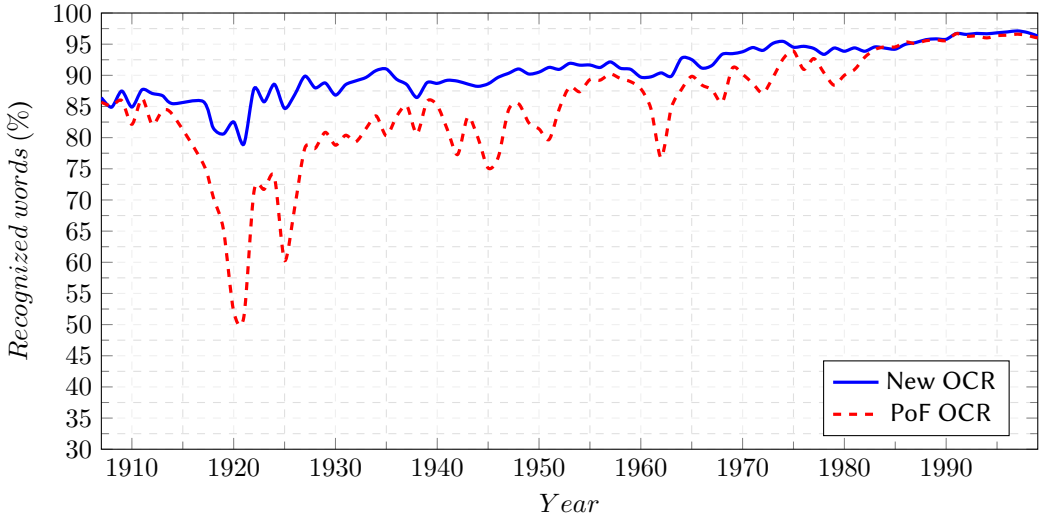



**Figure 2:** The percentage of recognized words with LAS tool using original PoF OCR results (red dashed line) and our new OCR (blue line) results.

Finally, 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 the set CSV tables.

2. **Corpus 1999–2014** From halfway 1999 to the end of 2014, the minutes were available also in already structured HTML form at the PoF's web pages[20]. The HTML documents were transformed into CSV tables.

3. **Corpus 2015–** The plenary session minutes from 2015– are available also based on a custom-made XML schema from the *Avoin eduskunta* API[21]. These XML documents were transformed into the CSV tables.

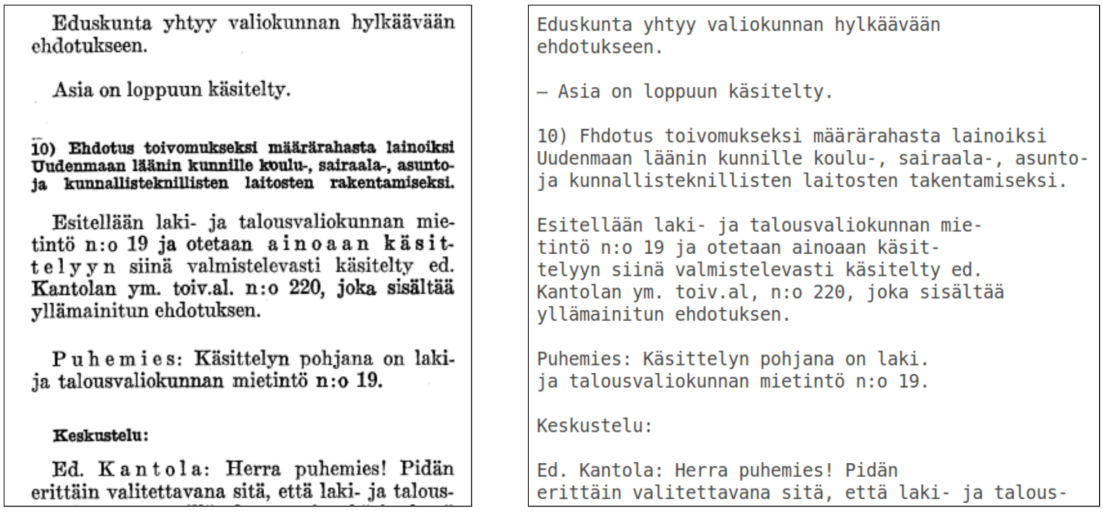
Each source format 1–3 differs in terms of the metadata included in the minutes. However, all formats contained the following core metadata elements about the session, speaker, and the speech: 1) Session data: session identifier, session date, session ending and starting times 2) Speaker data: last name, speaker's role/title 3) Speech data: speech content, speech type, related documents, and debate topic.

In the final speech CSV tables each row contains an individual speech with the content and metadata elements represented in columns.

Figure 3 shows an example of the 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: (debate/conversation)*). 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/-*

[20]Available at: https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx
[21]Open PoF API: https://avoindata.eduskunta.fi/#/fi/home

*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.



Eduskunta yhtyy valiokunnan hylkäävään ehdotukseen.

Asia on loppuun käsitelty.

**10) Ehdotus toivomukseksi määrärahasta lainoiksi Uudenmaan läänin kunnille koulu-, sairaala-, asunto- ja kunnallisteknillisten laitosten rakentamiseksi.**

Esitellään laki- ja talousvaliokunnan mietintö n:o 19 ja otetaan ainoaan käsittelyyn siinä valmistelevasti käsitelty ed. Kantolan ym. toiv.al. n:o 220, joka sisältää yllämainitun ehdotuksen.

Puhemies: Käsittelyn pohjana on laki- ja talousvaliokunnan mietintö n:o 19.

**Keskustelu:**

Ed. Kantola: Herra puhemies! Pidän erittäin valitettavana sitä, että laki- ja talous-

```
Eduskunta yhtyy valiokunnan hylkäävään
ehdotukseen.

— Asia on loppuun käsitelty.

10) Fhdotus toivomukseksi määrärahasta lainoiksi
Uudenmaan läänin kunnille koulu-, sairaala-, asunto-
ja kunnallisteknillisten laitosten takentamiseksi.

Esitellään laki- ja talousvaliokunnan mie-
tintö n:o 19 ja otetaan ainoaan käsit-
telyyn siinä valmistelevasti käsitelty ed.
Kantolan ym. toiv.al, n:o 220, joka sisältää
yllämainitun ehdotuksen.

Puhemies: Käsittelyn pohjana on laki.
ja talousvaliokunnan mietintö n:o 19.

Keskustelu:

Ed. Kantola: Herra puhemies! Pidän
erittäin valitettavana sitä, että laki- ja talous-
```

**Figure 3:** OCR example. On the left is a part of the original PDF-document; on the right is the same part with recognized text. [40]

The structure of the CSV tables 1907–1999 and the CSV tables based HTML-formatted minutes in 2000–2014 are fairly similar with over 20 metadata fields, such as speech identifies, session, data, start and end times, name of the speaker, his/her party and so on. Starting from 2015 the minutes are available as XML files; the corresponding CSV table format contains the following columns for metadata about speeches: party, topic, content, speech_type, status, version, link, lang, name_in_source, speaker_id, speech_start, speech_end, speech_status, and speech_version. More documentation about the data can be found in the Allas Store site.

### 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, text "... nostamiseksi [Arto Satosen välihuuto] hallitusohjelman ... " means that MP Arto Satonen made an interruption (shouted something) at this point of a fellow speakers's speech. 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 the PoF ("pöytäkirjatoimisto" in Finnish) [45]. It is not fully known what kind of changes in practice there have been at different times. These changes may have implications on 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 in different times is analyzed, this change in the 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, the speeches from 1925 are in the file `speeches_1925.csv`. However, occasionally there have been two parliamentary sessions referring to the same calendar year[22]. 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 at:

**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.

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

The XML TEI-based Parla-CLARIN [41] schema is an attempt to define a common XML-based annotation model for parliamentary debates on an international level.[23] For example, the Slovene parliamentary corpus siParl (1990–2018) has been encoded with the Parla-CLARIN schema [21]. Currently, the Parla-CLARIN schema is implemented in the Clarin ParlaMint project[24], which establishes a comparable and interoperable corpus of European parliamentary

[22] Due to the Government resigning prematurely and thus starting a new parliamentary session

[23] See: https://www.clarin.eu/blog/clarin-parlaformat-workshop

[24] https://github.com/clarin-eric/ParlaMint

corpora for comparative research. This format is a specialization of Parla-CLARIN extending it with, for example, linguistic and named entity mention annotations.

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 places in a 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 using a file system similar to that of the CVS tables:

**https://a3s.fi/parliamentsampo/speeches/xml/index.html**

The ParlaMint subcorpus is under validation and will appear later in the ParlaMint data repository[25].

### Publication as Linked Open Data

The LOD version of the speech data was created from the CSV tables, too [9, 40]. The latest corpus 2015– has been annotated semantically using Natural Language Processing (NLP) techniques as discussed in [46]:

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 [47, 48].

2. **Automatic keyword annotation**. Finnish NLP technology was applied also for annotating the speeches automatically using the YSO ontology[26] [49] of the National Library of Finland and the Annif automatic annotation tool[27] [50]. Ontology-based keywords facilitate semantic search and content-based analyses of the speeches. The data includes also keywords extracted using the traditional TF-IDF method.

3. **Automatic library classification**. The EKS subject headings[28] vocabulary of the Library of Parliament and Archive of Parliament was transformed into a SKOS[29] 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.

[25] See the current ParlaMint 2.1 version: http://hdl.handle.net/11356/1432
[26] https://finto.fi/yso/fi/
[27] https://annif.org/
[28] https://www.eduskunta.fi/kirjasto/EKS/
[29] Simple Knowledge Organization System: https://www.w3.org/TR/skos-reference/

The NLP-based annotations have been published as part of the ParliamentSampo RDF Turtle data dump in Zenodo.org[30] and as linked open data on the Linked Data Finland platform[31].

### Data Models for Speeches and Their Annotations

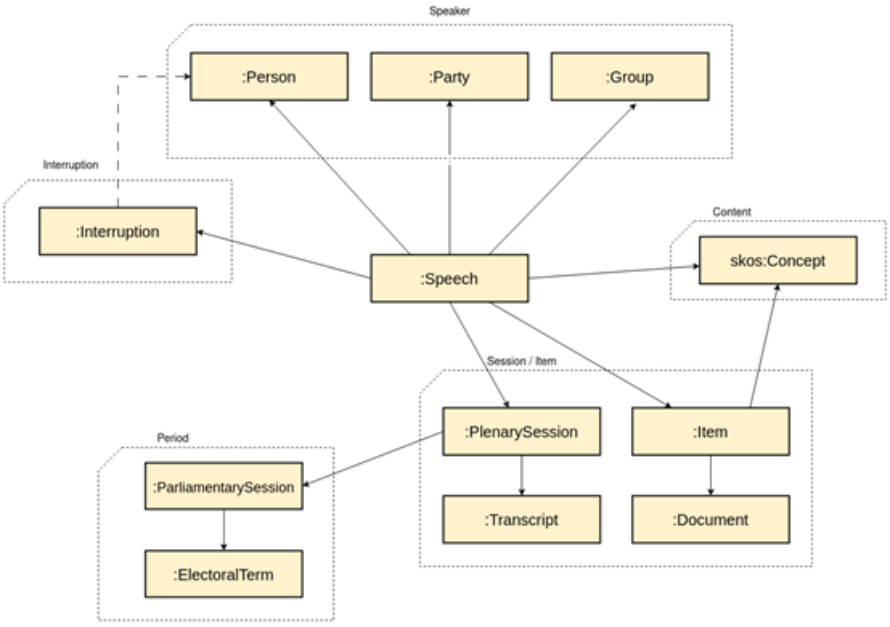



**Figure 4:** Data model for speech data in the default namespace https://ldf.fi/schema/semparl/

The data model of speech data is depicted in Figure 4; additional documentation can be found in [9, 40]. 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 annoations can be seen in Figure 5. More documentation about the these data models can be found using the namespace URL in a browser.

## 4. Using the ParliamentSampo Data

This section discusses briefly different way of using the ParliamentSampo infrastructure described above.

### 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 services presented above for local use, and then apply one's favourite tools for data analysis,

[30] https://doi.org/10.5281/zenodo.7636420
[31] https://www.ldf.fi/dataset/semparl

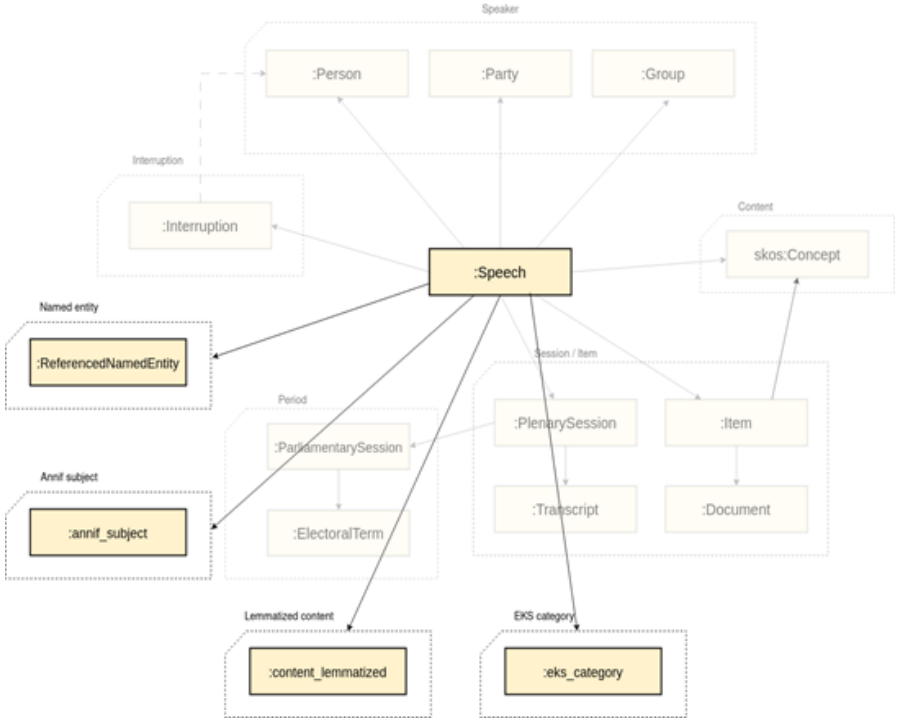



**Figure 5:** Data model for the linguistic annotations of speech data 2015– in the default namespace https://ldf.fi/schema/semparl/, with the related speech data model in the background

such as spreadsheets, R[32] environment for statistical analysis, or Gephi[33] 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 for linked data on one's own computer, for example Fuseki[34], which is also used in the LDF.fi service. The materials in the LDF.fi service are published using container technology (i.e., Docker[35]), 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 [32]. 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

[32] https://www.r-project.org
[33] https://gephi.org
[34] https://jena.apache.org/documentation/fuseki2/
[35] https://www.docker.com

language from protecting the nature to issues of climate change.

## 4.2. Querying the Endpoint and Studying Results

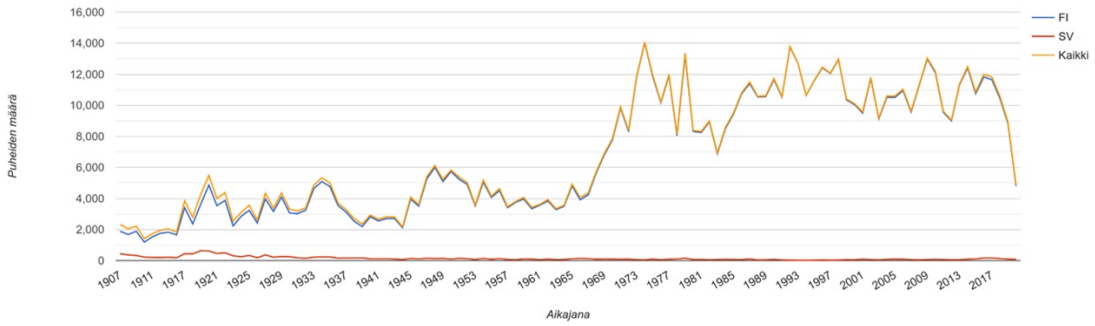



**Figure 6:** Number of speeches in different languages (y-axis) on the timeline (x-axis).

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, Figure 6 shows a visualization of the number of Finnish (FI), Swedish (SV) and all (Kaikki) speeches (y-axis) in the S-KG graph on a timeline from 1907 to 2021 (x-axis). Before the WW2, there have been more speeches in Swedish than today, but the number remains very small. The graphic was created using the YASGUI editor[36] [51], which can be used to edit SPARQL queries, target them to an online SPARQL endpoint, and to show the results using pre-implemented visualizations.

## 4.3. Data-analysis by Scripting

The PoF data can be examined computationally, for example, using Python scripting and Jupyter notebooks in the Google Colab[37] 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 analysis of using Google Colab is presented in Figure 7. It presents the yearly (x-axis) average lengths (y-axis) of speeches of all speakers (Kaikki), male speakers (Mies), and female speakers (Nainen), as well as the raising proportion of speeches by female speakers (Naisten osuus).

## 4.4. Using the ParliamentSampo Portal

The ParliamentSampo portal, based on the Sampo model [6] and the Sampo-UI framework [7], demonstrates how the SPARQL data service can be used for developing applications for DH

[36] https://yasgui.triply.cc
[37] https://colab.research.google.com

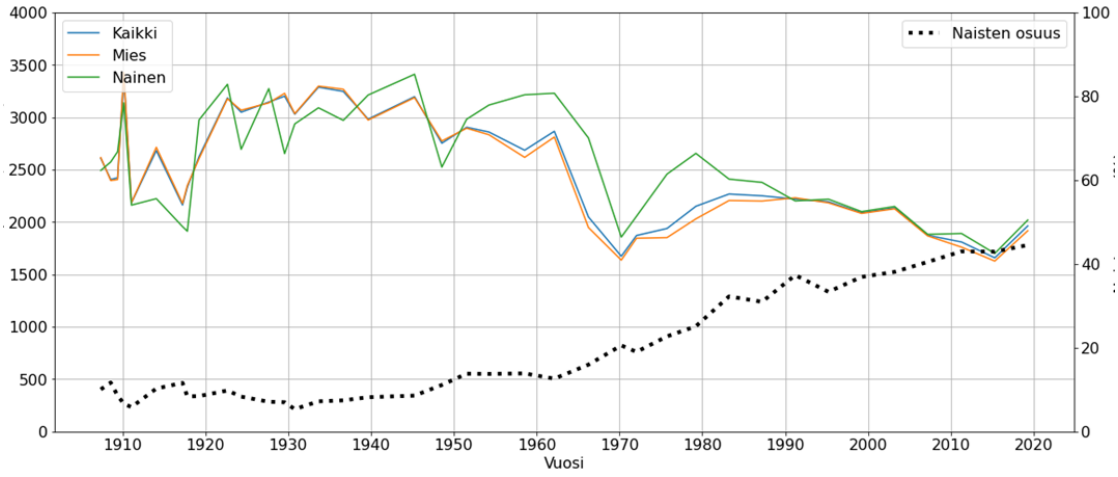



**Figure 7:** Average annual lengths of all (kaikki), male (mies), and female (nainen) speakers, and the raising proportion of speeches by female speakers (naisten osuus).

research. In the portal, the data can be filtered using faceted search [52] based on ontologies, and the results can then be analyzed with the help of seamlessly integrated visualizations and data analytic tools. The data can be accessed along application perspectives for studying 1) speeches of different times and 2) MPs and other speakers.

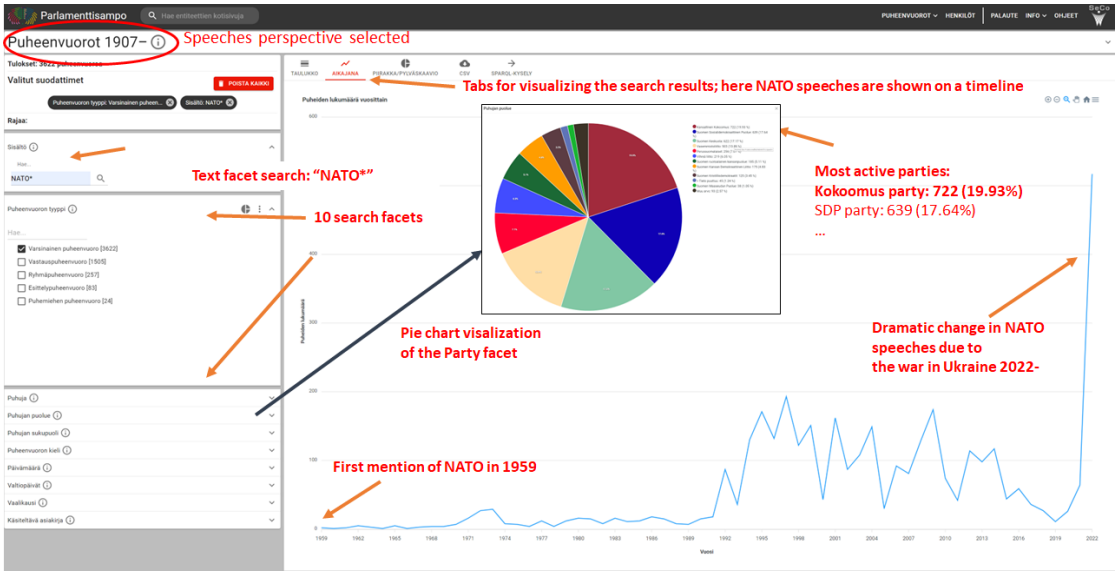



**Figure 8:** Using faceted search to filter out and analyze speeches about NATO.

For example, in Figure 8, the user has selected the Speeches perspective with facets Content,

Speaker, Party, (Speech) Type, and others on the left. The search result, i.e., the speeches found, is shown by default in tabular form on the right, but the results can also be visualized in other forms by selecting one of the five tabs: here the timeline visualization is used. The user has written a query "NATO*" in the Content text facet, the speech type is set to regular speeches, and then 3622 regular speeches that mention the word "NATO" in its various inflectional forms have been filtered into the search result starting from 1959. In addition, by clicking on the pie chart visualization button on the Party facet, the distribution of NATO speeches in terms of parties is shown: the most active party with 722 speeches has been the right wing National Coalition Party Kokoomus.

## 5. Discussion

The speech datasets of ParliamentSampo presented in this paper make it possible to find and study 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-undestandable" 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 harmonized enriched open data with data 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.

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 that of the 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., [53, 54]) and the language of politics and its long-term changes can be studied (e.g., [55, 56, 57, 58, 59, 60, 38]). 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., [61, 47, 48]).

Planned future development of ParliamentSampo includes using and extending the system in parliamentary research studies, correcting the historical data based on user feedback that is collected, e.g., using the portal, validating the data using ShEx shape expressions[38], and maintaining the data services as part of the national FIN-CLARIA/DARIAH-FI research infrastructure program[39].

**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, are acknowledged regarding the use cases and research on parliamentary culture. 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

[38] https://shex.io/

[39] https://seco.cs.aalto.fi/projects/fin-clariah/

of Finland in the projects Semantic Parliament[40] and FIN-CLARIAH[41], by CLARIN.eu in the ParlaMint II project[42]. OUr work is also related to the EU project InTaVia[43] and the EU COST action Nexus Linguarum[44] on linguistic linked data data resources and analysis. Thanks to Finnish Cultural Foundation for the Eminentia Grant of the first author. The project uses the computing resources of the CSC – IT Center for Science.

## References

[1] C. Benoît, O. Rozenberg (Eds.), Handbook of Parliamentary Studies: Interdisciplinary Approaches to Legislatures, Edward Elgar Publishing, 2020. doi:10.4337/9781789906516.

[2] M. Hidén, H. Honka-Hallila, Miten eduskunta toimii, Edita Publishing, Helsinki, 2006.

[3] 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.

[4] E. Hyvönen, P. Leskinen, L. Sinikallio, S. Drobac, R. Leal, M. La Mela, J. Tuominen, H. Poikkimäki, H. Rantala, ParliamentSampo infrastructure for publishing the plenary speeches and networks of politicians of the Parliament of Finland as open data services, Paper presented at the publication event of the ParliamentSampo infrastructure, University of Helsinki, February 14th, 2023. URL: https://seco.cs.aalto.fi/publications/2023/hyvonen-et-al-ps-data-2023.pdf.

[5] E. Hyvönen, L. Sinikallio, P. Leskinen, M. La Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, Finnish parliament on the semantic web: Using ParliamentSampo 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] E. Hyvönen, Digital humanities on the semantic web: Sampo model and portal series, Semantic Web – Interoperability, Usability, Applicability 14 (2022) 729–744. doi:10.3233/SW-190386.

[7] 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.

[8] 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.

[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

[40] https://seco.cs.aalto.fi/projects/semparl/
[41] https://seco.cs.aalto.fi/projects/fin-clariah/
[42] https://www.clarin.eu/parlamint
[43] https://intavia.eu
[44] https://nexuslinguarum.eu

Dagstuhl- Leibniz-Zentrum fur 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: Further with Knowledge Graphs. Proceedings of the 17th International Conference on Semantic Systems, 6-9 September 2021, Amsterdam, The Netherlands, IOS Press, 2021, pp. 255–269. doi:10.3233/SSW210049.

[11] E. Hyvönen, L. Sinikallio, P. Leskinen, S. Drobac, J. Tuominen, K. Elo, M. La Mela, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, Parlamenttisampo: eduskunnan aineistojen linkitetyn avoimen datan palvelu ja sen käyttömahdollisuudet, Informaatiotutkimus 40 (2021). doi:10.23978/inf.107899.

[12] 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, in: Digital Humanities in the Nordic and Baltic Countries, 7th Conference, CEUR Workshop Proceedings, 2023. URL: https://seco.cs.aalto.fi/publications/2022/drobac-et-al-ocr-2022.pdf, in press.

[13] E. Hyvönen, H. Rantala, P. Leskinen, Integrating faceted search with data analytic tools in the user interface of ParliamentSampo – Parliament of Finland on the Semantic Web, in: Proceedings of ESWC 2023, Poster and Demo Papers, Springer, 2023. URL: https://seco.cs.aalto.fi/publications/2023/hyvonen-et-al-ps-eswc-2023.pdf, paper submitted for peer review.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] U. Bojārs, R. Darģis, 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.

[19] R. Bleier, F. Zeilinger, G. Vogeler, From early modern deliberation to the semantic web: Annotating communications in the records of the Imperial Diet of 1576, in: M. La Mela, F. Norén, E. Hyvönen (Eds.), Proceedings of the Digital Parliamentary Data in Action (DiPaDA 2022) Workshop co-located with 6th Digital Humanities in the Nordic and Baltic Countries Conference (DHNB 2022), volume 3133, CEUR WS, 2022, pp. 86–100. URL: http://ceur-ws.org/Vol-3133/paper06.pdf.

[20] M. Ogrodniczuk, P. Osenova, T. Erjavec, D. Fišer, N. Ljubešic, Çagrı Çöltekin, M. Kopp, K. Meden, ParlaMint II: The show must go on, in: 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, pp. 1–6. URL: https://aclanthology.org/2022.parlaclarin-1.1.pdf.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] J. Jarlbrink, F. Norén, The rise and fall of 'propaganda' as a positive concept: a digital reading of Swedish parliamentary records, 1867–2019, Scandinavian Journal of History (2022) e1–e21. doi:10.1080/03468755.2022.2134202.

[30] 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.

[31] K. Kettunen, M. La Mela, Semantic tagging and the nordic tradition of everyman's rights, Digital Scholarship in the Humanities 37 (2021). doi:10.1093/llc/fqab052.

[32] K. Elo, J. Karimäki, Luonnonsuojelusta ilmastopolitiikkaan: Ympäristöpoliittisen käsitteistön muutos parlamenttipuheessa 1960–2020, Politiikka 63 (2021). URL: https://journal.fi/politiikka/article/view/109690. doi:10.37452/politiikka.109690.

[33] 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.

[34] A. Martínez Arranz, S. T. Zech, M. Bonotti, Political Parties and Civility in Parliament: The Case of Australia from 1901 to 2020, Parliamentary Affairs (2023). doi:10.1093/pa/gsad008, gsad008.

[35] 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.

[36] 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.

[37] S. Simola, A century of partisanship in Finnish political speech, 2020. URL: https://sites.google.com/site/sallasimolaecon/home/research.

[38] 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.

[39] 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.

[40] L. Sinikallio, Eduskunnan täysistuntojen pöytäkirjojen muuntaminen semanttiseksi dataksi ja julkaiseminen verkkopalveluna, Master's thesis, University of Helsinki, Department of Computer Science, 2022. URL: http://urn.fi/URN:NBN:fi:hulib-202204201707.

[41] T. Erjavec, M. Ogrodniczuk, P. Osenova, et al., The ParlaMint corpora of parliamentary proceedings, Lang Resources & Evaluation 57 (2022) 415–448. doi:10.1007/s10579-021-09574-0.

[42] 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. doi:10.1007/978-3-319-11955-7_24.

[43] E. Lapponi, 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. doi:10.1007/s10579-018-9411-5.

[44] E. Mäkelä, LAS: an integrated language analysis tool for multiple languages., J. Open Source Software 1 (2016) 35. doi:10.21105/joss.00035.

[45] Kirjo – kirjaamisohjeet, Eduskunnan kanslia, Helsinki, Finland, 2021. Guidelines for recording minutes of plenary sessions at Parliament of Finland.

[46] 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, international Workshop on Knowledge Graph Generation from Text (TEXT2KG 2022).

[47] 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. URL: https://seco.cs.aalto.fi/publications/2022/poikkimaki-et-al-2022.pdf, accepted.

[48] H. Poikkimäki, Eduskunnan täysistuntojen puheenvuorojen henkilömainintoihin perustuvien verkostoiden analyysi, Master's thesis, Aalto University, Department of Computer Science, 2023. URL: https://seco.cs.aalto.fi/publications/2023/poikkimaki-msc-2023.pdf.

[49] K. Seppälä, E. Hyvönen, Asiasanaston muuttaminen ontologiaksi. Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista, National Library, Plans, Reports, Guides, 2014. URL: https://www.doria.fi/handle/10024/96825.

[50] O. Suominen, Annif: DIY automated subject indexing using multiple algorithms, LIBER Quarterly 29 (2019) 1–25. doi:10.18352/lq.10285.

[51] L. Rietveld, R. Hoekstra, The YASGUI family of SPARQL clients, Semantic Web – Interoperability, Usability, Applicability 8 (2017) 373–383. doi:10.3233/SW-150197.

[52] Y. Tzitzikas, N. Manolis, P. Papadakos, Faceted exploration of RDF/S datasets: a survey, Journal of Intelligent Information Systems 48 (2017) 329–364.

[53] D. Mimno, Topic Regression, Ph.D. thesis, University of Massachusetts Amherst, 2012. URL: https://scholarworks.umass.edu/open_access_dissertations/520.

[54] 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.

[55] 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.

[56] 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.

[57] S. Purhonen, A. Toikka, "Big Datan" haaste ja uudet laskennaliset tekstiaineistojen analyysimenetelmät: esimerkkitapauksena aihemallianalyysi tasavallan presidenttien uudenvuodenpuheista 1935–2015, Sosiologia 53 (2016) 6–27. URL: http://elektra.helsinki.fi/se/s/0038-1640/53/1/bigdatan.pdf.

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

[59] 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.

[60] J. B. Mountford, Topic modeling the red pill, Social Sciences 7 (2018). doi:10.3390/socsci7030042.

[61] 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.