
Eero Hyvönen¹², Petri Leskinen¹², Laura Sinikallio²¹, Matti La Mela², Jouni Tuominen¹², Kimmo Elo³, Senka Drobac²¹, Mikko Koho¹², Esko Ikkala¹, Minna Tamper¹², Rafael Leal¹² and Joonas Kesäniemi¹

¹Semantic Computing Research Group (SeCo), Aalto University, Finland
²Helsinki Centre for Digital Humanities (HELDIG), University of Helsinki, Finland
³Centre for Parliamentary Studies, University of Turku, Finland

Abstract
This paper introduces the system ParliamentSampo – Parliament of Finland on the Semantic Web, a Linked Open Data (LOD) service, data infrastructure, and semantic portal for studying Finnish political culture, language, and networks of the Members of Parliament (MP). The article presents the vision behind the system, the LOD service, and explores the possibilities to utilize it in research and application development. A knowledge graph of linked data has been created based on more than 950,000 speeches in all plenary sessions of the parliament in 1907—2021; the data is also available in XML format, utilizing the new international Parla-CLARIN format. For the first time, the entire time series of the Finnish parliamentary speeches has been converted into data and a data service in a unified format. In addition, the speeches have been interlinked with another knowledge graph created from the database of Members of Parliament (MP) and enriched from other data sources into a broader ontology-based data service. The paper shows how the LOD service SPARQL endpoint can be used to research parliamentary culture, the use of political language, and networks of politicians through data analysis. The service endpoint can also be used to develop applications for different user groups, such as the PARLIAMENTSampo semantic portal, introduced in the paper, too.

Keywords
parliamentary studies, semantic portals, linked data, digital humanities

1. Introduction

The main tasks of parliaments are to enact new laws, oversee the work of the government, and decide on the state budget; how the parliament works in Finland is documented in [1]. Parliamentary data are used in many areas of research [2], as it provides a wealth of information on the state and functioning of democratic systems, political life and, more generally, language and culture. The most prominent part of the work of parliaments is the public plenary sessions, in which Members of Parliament (MP) discuss and vote on issues on the agenda and other topics that arise. Parliaments draw up minutes of plenary sessions and make both the minutes and the
documents on which they are based on available to the public. Openness and transparency in the work of parliaments is important so that voters, the media, researchers, and also parliaments themselves can look at the decision-making stages and the views and actions expressed by parliamentarians in their work as legislators.

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

We test and demonstrate the argument in DH research on parliamentary culture and language [2] by presenting the PARLIAMENTSAMPO system, a Linked Open Data (LOD) corpus and data service of Finnish parliamentary data and a semantic portal on top of it\(^2\). The paper first reviews related research and clarifies the vision of publishing and using parliamentary data on the SW. First results obtained in developing and using the PARLIAMENTSAMPO are presented. In conclusion, using parliamentary data in research is considered on a more general level.

2. Related Work on Parliamentary Data

Lots of parliamentary materials have been digitized in recent decades, arguably only second to newspapers [9]. For example, the Royal Library of Sweden has digitized Swedish printed parliamentary documents from 1521 to 1970. This collection\(^3\) is supplemented by the parliament’s own digital materials and, e.g., by the Westac research project\(^4\) at the Umeå University. Digitization has improved the accessibility and usability of parliamentary materials for both the public and the research community. Websites have been created that make it easy for users to browse and download materials. Examples include the website of the Lipad project\(^5\) [10] that digitized Canadian parliamentary materials, and the portal Italian House of Representatives\(^6\) that comprehensively presents the history of the Italian parliament in 1848–2018.

Several parliamentary corpora have been formed from the minutes of the plenary debates, which make it possible to study the content of the speeches and their language; see, e.g., [11] and the CLARIN list of parliamentary corpora\(^7\). The TEI-based Parla-CLARIN scheme\(^8\) for session minutes has been developed within the CLARIN infrastructure, providing a common way to

---

1 https://www.w3.org/standards/semanticweb/
2 See the project homepage for more details, videos, and publications: https://seco.cs.aalto.fi/projects/semparl/en/
3 http://data.riksdagen.se
4 https://www.westac.se
5 https://lipad.ca
6 https://storia.camera.it
7 https://www.clarin.eu/resource-families/parliamentary-corpora
8 https://github.com/clarin-eric/parla-clarin
represent the corpora [12]. The related ParlaMint project brings together Parla-CLARIN-based national corpora. Parliamentary materials have also been transformed into the form of LD when creating the LinkedEP [3] system on the European Parliament’s data, the Italian Parliament[10], and the LinkedSaeima for the Latvian parliament [4].

The materials of the Parliament of Finland (PoF) have been digitized in various contexts but are difficult to use, as they have been produced separately from different periods and stored in different formats [13]. The usability of the materials is also hampered by their varying quality and lack of descriptive data [14]. Language corpora have been published on parliamentary debates, such as the Parliamentary Corpus of FIN-CLARIN’s Language Bank [15] which covers the years 2008–2016. It contains the speeches in a linguistically annotated form and also synchronized links to original plenary session videos [16]. The Voices of Democracy project has produced a research corpus that includes plenary minutes in 1980–2018 annotated grammatically as well as interviews of veteran MPs conducted by the PoF after 1988 [9]. The minutes of the parliamentary debates from 1991 to 2015 can also be found in the International Harvard Parlspeech Corpus [17], but we have identified gaps in the coverage in this corpus.

Digitized parliamentary materials offer a wide range of perspectives on different research topics and have been used in a variety of fields, such as linguistics, political science, media studies, economics, and history. The most important research material are the debates in the parliaments, through which one can study at the language and its changes itself as well as the underlying societal phenomena at large [18]. Metadata makes it possible to structure the speeches, for example, between parties, gender, or professional groups. Luke Blaxill and Kaspar Beelen [19] have examined the content of women’s parliamentary speeches, as well as the role of gender in the speeches of MPs in the British Parliament. Parliamentary debates have been used in thematic or conceptual analyses (cf., e.g., [20, 21, 22, 23, 24] and to study the language or the opinions of the parties or MPs (e.g., [25, 26]). Parliamentary debates have also been used in translation studies using, for example, the EuroParl Corpus of European Parliament debates.

The digitized material of the Finnish Parliament has been utilized to some extent in digital humanities and social scientific research. Matti La Mela [14], also Kettunen and La Mela [24], have studied the history of the concept of Everyman’s right, a Nordic right of public access to nature, with the digitized minutes of the Parliament, and examined their quality in PDF format. The digitized minutes have been utilized in the development of language technology methods, in this case the Finnish Semantic Tagger [24]. Similarly, Andrushschenko et al. [9] have used their grammatically structured corpus and a search tool to organize and analyze parliamentary debates in various research cases. Salla Simola [27] has examined the differences in political speech between parties throughout the parliamentary period 1907–2018, for which she compiled a separate research dataset combining the debates and the speaker data. Kimmo Makkonen and Petri Loukasmäki [28] have studied the plenary speeches given in Parliament in 1999–2014 and their content by using topic modeling. FIN-CLARIN’s Parliamentary Corpus has been used, for example, by Ella Lillqvist et al. [29] in their study on debates about public debt. Previous search applications for Finnish parliamentary speech data are based mostly on traditional text.

---

9https://www.clarin.eu/content/parlamint-towards-comparable-parliamentary-corpora
10http://data.camera.it
11http://korp.csc.fi
12https://www.statmt.org/europarl/
search. Data analysis tools to examine the results are few, such as the concordance analysis of the Language Bank Korp, where the words found are visualized in their textual contexts and show some statistics of words occurrences in the search results. These applications cover only a small part of the entire time series of the Finnish parliamentary speeches.

3. ParliamentSampo Vision

The vision of the Semantic Parliament project [30] is to develop and implement in the living laboratory environment the model shown in Fig. 1 for publishing and utilizing parliamentary materials as LOD on the SW. The work focuses on two core datasets:

1. **Minutes of Parliamentary Sessions** All Finnish parliamentary debates, totalling over 950,000 speeches and covering the existence of the PoF 1907–2021, have been transformed into a 1) Linked Data knowledge graph and into 2) Parla-CLARIN XML form. [13]

2. **Members of Parliament Data** A proposographical knowledge graph representing biographical data about the politicians during the same time period (1907–2021). [31]

![Figure 1: Vision: Linked Open Data publishing model of PARLIAMENTSAMPO.](http://seco.cs.aalto.fi/projects/lawlod/)
the PoF and 2) a set of related vocabularies and ontologies describing, for example, (historical) places\(^\text{14}\), professions \(^\text{35}\), people, and organizations.

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

### 4. Core Knowledge Graphs in the LOD Service

**Speech Graph S-KG** The knowledge graph of parliamentary speeches (S-KG) (cf. Fig. 1), contains speeches collected from all the minutes of the plenary sessions of the PoF since 1907. The S-KG was compiled from several initial formats: 1) From 1907 until the middle of 1999, the minutes are available only as scanned images embedded in PDF documents. This material was OCRed with minor manual corrections made. 3) From mid-1999 to the end of 2014, the material was available in HTML format at the Parliament’s website\(^\text{15}\). 3) From the 2015 onwards, the minutes are available through the Finnish Parliament Open Data API\(^\text{16}\) in custom XML form.

The data model of S-KG and the data transformation process are described in detail in \(^\text{13}\).

**Protopographical Graph P-KG** The S-KG was interlinked to the MPs protopographical knowledge graph P-KG (cf. Fig. 1). For example, speakers and the parties they represent are resources with URI identifiers described in the P-KG graph. The data publication about MPs is a knowledge graph (P-KG) covering all MPs who have worked in Finland \(^\text{31}\). At its core is an RDF conversion of XML-formatted data about MPs downloaded from the Open Data service\(^\text{17}\) of PoF. In addition to basic information such as times and places of birth and death, the data includes more detailed information about people’s life events, such as studying, working life, political career, or their written publications. The Parliament’s open data source has been supplemented with information extracted from the Government’s website\(^\text{18}\) and Wikidata: in addition to MPs, 200 other people of significant political history, such as presidents, ministers and ombudsmen, have been added to the knowledge graph. For example, Mauno Koivisto has served as President and Prime Minister but never as an MP.

The data model of P-KG, based on the CRM Bio extension \(^\text{36}\) of CIDOC CRM\(^\text{19}\), as well as the data transformation process are described in more detail in \(^\text{31}\).

### 5. Using the PARLIAMENTSAMPO Data Service

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

\(^{14}\)https://seco.cs.aalto.fi/projects/histoplaces/
\(^{15}\)https://www.eduskunta.fi/FI/taynistunto/Sivut/Taynistuntojen-poytakirjat.aspx
\(^{16}\)https://avoindata.eduskunta.fi/#/fi/home
\(^{17}\)https://avoindata.eduskunta.fi/#/fi/dbsearch
\(^{18}\)https://valtioneuvosto.fi/hallitukset-ja-ministerit
\(^{19}\)https://cidoc-crm.org
Exporting the Data for External Use: A simple way for a researcher to use PARLIAMENTSAMPO data is to download data from the data service for local use and then apply one’s favourite tools for data analysis, such as spreadsheets, R\textsuperscript{20} environment for statistical analysis, or Gephi\textsuperscript{21} for network analysis. For filtering out subsets of interest in the big data, SPARQL querying can be used in flexible ways. It is also possible to install a local SPARQL server environment for linked data on one’s own computer, for example Fuseki\textsuperscript{22}, which is also used in the LDF.fi service. The materials in the LDF.fi service are published using container technology (i.e., Docker\textsuperscript{23}), which means that installing the data, the server, and possible versioned software packages is automatic and effortless.

An example using PARLIAMENTSAMPO data externally is reported in [18]. For this case study in political science, the Parla-CLARIN version was downloaded and a subset of the speeches 1960–2020 was filtered out and analyzed further using custom XML-based tools. The authors studied how the language used in discussing environmental politics has evolved in Finland in the speeches of different parties. Eleven central environmental terms were selected from a thesaurus used by the PoF library, speeches where these terms were used were then extracted, and various quantitative analyses based on them were presented and compared with the strategies 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 2: Number of speeches in different languages on the timeline.

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. 2 shows a visualization of the number of speeches Y-axis) in the S-KG graph by language on a timeline from 1907 to 2020 (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

\textsuperscript{20}https://www.r-project.org
\textsuperscript{21}https://gephi.org
\textsuperscript{22}https://jena.apache.org/documentation/fuseki2/
\textsuperscript{23}https://www.docker.com
speeches in Swedish (‘SV’ in the figure) than today, but the number remains very small. The
graphic was created using the YASGUI editor24 [37], which can be used to edit SPARQL queries,
target them to an online SPARQL endpoint, and to show the results using pre-implemented
visualizations. The SPARQL query used to generate Fig. 2 is shown below:

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

SELECT ?year (COUNT(?fin) as ?FI) (COUNT(?swe) as ?SV)
(COUNT(?document-URI) as ?ALL ) WHERE {
BIND(STR(year(?dateTime)) as ?year)
{ 
BIND( <http://id.loc.gov/vocabulary/iso639-2/swe> as ?swe)
} UNION {
BIND( <http://id.loc.gov/vocabulary/iso639-2/fin> as ?fin)
}
GROUP BY ?year ORDER BY ASC(?year)
```

SPARQL is an expressive and flexible way to retrieve information from graphical data, and it
is suitable for use by digital humanities researchers. The SPARQL query above first introduces
the namespaces used (PREFIX). In the next SELECT query, all speeches and their languages
are retrieved using a graph formed by variables starting with ?, which is fitted to the end point
graph in all possible ways. The answer is a table of possible value assignments for the variables.
The results are finally classified (GROUP BY) into groups according to language, sorted by year
(ORDER BY), and finally it is summed up (COUNT) how many speeches there are in Finnish,
Swedish, and in total. In the visualization, the variable ?year forms the X-axis and the Y-axis
presents the annual number of speeches in different languages. Speech language recognition has
been done automatically, and sometimes OCR errors, for example, can make language recognition
difficult. Speeches that do not have a language code identified are excluded automatically from
the query result.

**Data-analysis by Scripting** The PoF data can be examined computationally, for example,
using Python scripting and Jupyter notebooks in the Google Colab25 environment. Then one can
use the simple HTTP protocol to perform SPARQL queries and then analyze and visualize query
results using tools provided by the programming environment used, such as Python libraries.

For example, Fig. 3 shows the starting age of persons elected as MPs for the first time each
year [31]. In the figure, a blue solid line shows the age of all MPs, and the age of women is
shown in red. It can be seen from the graph that the starting age has remained almost constant
throughout the parliamentary activities, but on the other hand, since 1980, women have been
younger than men for some time when they started as MPs. The relative proportion of women is
shown by a black dotted line. Before the 1960s, the proportion remained at an average of 10%,
but has since risen to 30–50%. The graphics in the image were implemented in Google Colab
using ready-made Python libraries for data analysis.

---

24https://yasgui.triply.cc
25https://colab.research.google.com
Figure 4 shows a similarly formed tabular visualization of the correlation between the parties and the occupations of the MPs. Here only the most common of the parties and occupations over the entire history of PoF are considered. The parties are presented in the horizontal rows of the table and the number of representatives of each profession is indicated in the vertical row corresponding to the occupation. The matrix shows, for example, that in the Centre Party, the National Coalition Party, and the Swedish People’s Party the most common occupation is Farmer. On the other hand, Entrepreneur has been the most common occupations with the Finns Party.

Using the PARLIAMENTSAMPO Portal The PARLIAMENTSAMPO portal, based on the Sampo model [38] and the Sampo-UI framework [39], demonstrates how the FinnParla data service can be used for developing applications for DH research. In the portal, the data can be filtered using faceted search [40] based on ontologies, and the results can then be analyzed with the help of seamlessly integrated visualization and data analysis tools.

The data can be accessed along two application views for filtering and analysing 1) speeches and 2) MPs. For example, in Fig. 5, the user has selected the Plenary Speeches view, which shows the search facets Content, Speaker, Party, (Speech) Type, Language, and Date on the left. The search result, i.e., the speeches found, is shown by default in tabular form on the right. The user has written a query “suomettum*” in the Content text facet, in which case only speeches that contain the word “suomettuminen” (Finlandization) in its various forms have been filtered into the search result, as the wildcard “*” matches any string. The result in this case is 174 speeches (that are paged with 10 results per page). The user has also limited the result on the Date facet to speeches given since June 4, 1945, when Parliament began to convene after World War II. In faceted search, the filtering selections can be made flexibly in any order, and the search engine calculates a hit count for each subsequent facet selection, which tells how many results would be obtained in the result set if the selection in question is made next. For example, in the Speaker facet, a click on “Junnila, Tuure (1910-199)” selects Tuure Junnila’s seven speeches that mention Finlandization. The selection facets are created automatically using the parliamentary
ontology and knowledge graphs of the FinParla data. The hit count allows the user to be directed to selections that do not lead to dead ends where the result set is empty. In addition, the hit numbers provide an opportunity to investigate at the result set statistically for different facets. For example, clicking on the pie symbol in the Speaker facet opens the pie pattern which shows how many different speakers mention Finlandization in their different speeches. The winners are Mr. Georg C. Ehrnrooth (21 speeches) and Mr. Ben Zyskowicz (18 speeches).

In accordance with the Sampo model, a number of pre-implemented data analysis tools and visualizations, similar to those shown in the figures above, can be integrated into the application perspectives of PARLIAMENTSAMPO. In the future, the tools and visualizations can be found alongside the table visualization in Fig. 5 on their own tabs in the same way as, for example, in the AcademySampo’s user interface [41]; the components of the Sampo-UI framework [39] are reused in the implementation of both portals. Through these tools and visualizations, the project explores the potential of Artificial Intelligence for knowledge discovery in digital humanities research [42]: how could PARLIAMENTSAMPO assist a researcher in solving research problems, in explaining solutions, and in finding new research problems?

Figure 4: Table of correlations between MP professions and parties [31]
6. Discussion

In the context of political research, the parliamentary speech is considered an important form of political communication and political struggle. A parliamentary speech is not just any speech, but has its own structure and its own rules, which at the same time reflect the general position of Parliament. In addition, a parliamentary speech is an instrument of political struggle to expose competing goals, challenge the views of an adversary, and unlock deadlocked settings. Thus, a speech in Parliament is always also a political act, in which the words used are the weapons of political decision-making and which not only tell about the issues under discussion, but also reveal the different positions, values, and points of view of the speakers. [43]

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

The Semantic Parliament project is an important step in utilizing the plenary debates in PoF as part of the field of humanities research. Although the materials have always been available to researchers manually and for some years also electronically digitized in PDF format, the machine-readable data corpus now being prepared and published as a data service, together
with the ParliamentSampo portal, will integrate parliamentary plenary debates and other open materials into the digital humanities and national information infrastructure. This means in practice, for example, the opportunity for political scientists, historians, and linguists to extract, model, analyze, and visualize parliamentary speech through exploratory research, using a vast body of data covering the entire period of the modern PoF since 1907.

The possibility of exploratory data analysis opens up completely new possibilities and perspectives for the study of parliamentary speech. In traditional close reading, the researcher is forced to delimit the material very strongly already at the collection stage, which usually happens through either temporal or thematic delimitation – that is, either by focusing on a limited period of time or on limited themes. Digital methods make it possible to examine the material without such limitations, which makes it possible to examine the material using, for example, fully automatic or semi-automatic classification methods. In this way, it may be possible to find, for example, new themes and topics that have been sidelined in research in the past (cf. e.g., [44, 45]).

On the other hand, distant reading and classification of data without strong presuppositions also allows for a critical examination of previous research results, when the themes/topics generated by distant reading can be compared with the results obtained by other methods [46].

Another example of the possibilities offered by data is research on the language of politics and its long-term change (e.g., [47, 48, 49, 50, 51, 52]). Parliamentary big data enables large-scale and systematic application of language technology methods. Although parliamentary speech is also linguistically its own special form of speech, parliamentary speech also lives in time and thus reflects both the wider linguistic development and the social atmosphere of discussion and word choices that occur in it [28]. At the same time, the extensive data offer an opportunity to study the change in language use, for example, whether the social debate climate is polarized or “brutalized”, as politicians and media actors have repeatedly suggested in recent years.

The third opportunity offered by parliamentary data relates to linking the use of language more broadly to other social contexts of language users, such as education, age, and social networks. Language can also be approached in policy research on the assumption that language always reflects the wider world of values and ideas of its user, as well as his or her social status and context. Discursive coalitions, which can be constructed based on the language use of the speakers, thus offer an interesting opportunity to detach oneself from the frame of reference set by, for example, the party background and to focus analytical attention on networks built through the use of language. In previous studies, this type of approach has been able to connect experts to different ideological positions by analyzing the content of their texts [53], which we think can be well applied to the classification of MPs.

A few examples have been highlighted above where the utilization of parliamentary data would seem to allow for significant new research openings in parliamentary research. However, in the spirit of exploratory data analysis, it is worth highlighting the as-yet-unknown possibilities that gradually emerge as researchers begin to outline new hypotheses and research questions by examining and analyzing data. The potential of large data sets is surprising in their potential, which on the one hand requires an open-minded attitude towards the data and on the other hand underscores the growing responsibility of researchers working in data analysis. When it is no longer possible for the researcher to know the material he or she is using thoroughly, he or she must know the phenomena that are the subject of the material thoroughly. Only in this way is it possible to assess which findings transmitted through excavation, analysis, modeling, or
Acknowledgements Our work is funded by the Academy of Finland and is also related to the EU project InTaVia and the EU COST action Nexus Linguarum. The project uses the computing resources of the CSC – IT Center for Science.

References


27https://nexuslinguarum.eu


