Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup

Laura Sinikallio
University of Helsinki, HELDIG Centre for Digital Humanities, SeCo Research Group, Finland

Senka Drobac
Aalto University, Department of Computer Science, SeCo Research Group, Finland

Minna Tamper
Aalto University, Department of Computer Science, SeCo Research Group, Finland

Rafael Leal
University of Helsinki, HELDIG Centre for Digital Humanities, SeCo Research Group, Finland

Mikko Koho
University of Helsinki, HELDIG Centre for Digital Humanities, SeCo Research Group, Finland

Jouni Tuominen
Aalto University and University of Helsinki (HELDIG), SeCo Research Group, Finland

Matti La Mela
University of Helsinki, HELDIG Centre for Digital Humanities, SeCo Research Group, Finland

Eero Hyvönen
Aalto University and University of Helsinki (HELDIG), SeCo Research Group, Finland

Abstract

This paper presents a knowledge graph created by transforming the plenary debates of the Parliament of Finland (1907–) into Linked Open Data (LOD). The data, totaling over 900 000 speeches, with automatically created semantic annotations and rich ontology-based metadata, are published in a Linked Open Data Service and are used via a SPARQL API and as data dumps. The speech data is part of larger LOD publication *FinnParla* that also includes prosopographical data about the politicians. The data is being used for studying parliamentary language and culture in Digital Humanities in several universities. To serve a wider variety of users, the entirety of this data was also produced using Parla-CLARIN markup. We present the first publication of all Finnish parliamentary debates as data. Technical novelties in our approach include the use of both Parla-CLARIN and an RDF schema developed for representing the speeches, integration of the data to a new Parliament of Finland Ontology for deeper data analyses, and enriching the data with a variety of external national and international data sources.

2012 ACM Subject Classification Information systems → Ontologies; Information systems → Resource Description Framework (RDF); Computing methodologies → Information extraction

Keywords and phrases Plenary debates, parliamentary data, Parla-CLARIN, Linked Open Data, Digital Humanities

Digital Object Identifier 10.4230/OASIcs.LDK.2021.12

1 Introduction

Semantic Parliament (SEMPARL)[1] is a consortium research project, which produces a linked open data and research infrastructure on Finnish parliamentary data, and develops novel semantic computing technologies to study parliamentary politics and political culture.

[1] https://seco.cs.aalto.fi/projects/semparl/en/

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch; Article No. 12; pp. 12:1–12:17
OpenAccess Series in Informatics
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

SEMPARL brings together researchers at the University of Helsinki, University of Turku, and Aalto University, with complementary, multi-disciplinary expertise in language technology, political and media research, and semantic computing and web technologies, respectively.

The project makes three major contributions. First, it responds to the demand for an easy to use and "intelligent" access to the newly digitized Finnish parliamentary data by providing the data as a national Linked Open Data (LOD) infrastructure and service for researchers, citizens, the government, and the media, and application developers. Second, the project studies long-term changes in the Finnish parliamentary and political culture and language. These use cases in political and language research are pioneering studies using the Finnish digital parliamentary data. Third, the new data service semantically enriches content in other related Finnish LOD services, such as LawSampo for Finnish legislation and case law [7] and BiographySampo for prosopographical data [6].

From a Linked Data production point of, two interlinked knowledge graphs (KG) are produced in SEMPARL: 1) A KG of all over 900 000 parliamentary debate speeches of the Parliament of Finland (PoF) (1907–present) to be called S-KG. 2) A prosopographical knowledge (P-KG) graph of the over 2600 Members of Parliament (MP), other people, and organizations related to the parliamentary speeches during the same period of time [16]. These KGs constitute together a larger data publication of PoF data called FinnParla. This paper presents the first graph S-KG and addresses the following more general research question: *How to represent and publish parliamentary speeches so that the data can be used easily for Digital Humanities research?*

In the following, we first present the problem of representing publishing, and using plenary debates as data for Digital Humanities research, and discuss related works and projects. After this, our original debate data, target data model, and the transformation process are described. The produced linked data has been published as a data service using the 7-star model of the Linked Data Finland platform [8]. As a demonstration of using the data service in Digital Humanities research, exemplary data-analyses are presented using YASGUI and Google Colab on top of the underlying SPARQL endpoint. In conclusion, contributions of the work are summarized, related works are discussed, and further research are outlined.

2 Related Work: Publishing Plenary Debates as Data

The Unicameral Parliament of Finland convened for the first time in 1907. The parliament has 200 members (MP), who are elected for four years. Since the first parliament of 1907, the elections are based on universal suffrage and both male and female MPs have been elected to all parliaments. In the Finnish parliament, the debates take place in the public plenary sessions. Since 1907, the Parliament has transcribed the speeches and published the printed plenary session minutes, which is a practice established already in the nineteenth-century Diet of Estates [20]. The minutes contain the matters considered, the decisions made, and every speech heard during the sessions. The wordings of the speeches are revised and improved for readability. [29, 20].

In the 1990s, the Parliament of Finland started to gradually publish parliamentary documents in digital form. It was only in 2018 that the Parliament completed the digitisation of the historical parliamentary documents of 1907–1999 and opened a new version of their data service [10]. This open data and the data service of the Parliament, however, has weaknesses concerning the data and its usability due to the heterogeneous data formats and different ways of access. For example, the historical minutes contain only the text recognised from the image files, and have no metadata concerning the structure of the minutes or their content, which limits the research to bag-of-words approaches [14].

There are also annotated corpora produced of the Finnish Parliamentary debates, which cover the recent decades. FIN-CLARIN has a curated corpus of the debates in 2008–2016 [3]. These include linguistic annotation, metadata about the speakers and the speeches are linked to the actual video recordings of the plenary sessions. Moreover, there is the multilingual Parlspeech parliamentary corpus [21], which includes also the plenary debates of the Finnish parliament in 1991–2015. This data, however, has quality problems. It has been created from the PDF files of the Parliament website of the time, but not all the speeches can be found in the data when we compare it with the complete minutes.

Several projects have transformed parliamentary debates into structured data or produced annotated parliamentary debate corpora. Regarding the former, the projects have foremost concerned the digitisation of the parliamentary debates and their enrichment with political or biographical metadata. These data have been transformed both to XML and RDF format[2]. In the Lipad project, the Canadian Hansard from 1901 to present was transformed into linked XML structured data [1]. As in our case, the process included both the OCR and the parsing of the historical documents and more straight-forward conversion of the recent SQL parliamentary data. The major example of parliamentary data in RDF is the Linked EP project, where the data of the European parliament 1999–2017 was transformed into RDF format and enriched with biographical information [28]. The RDF standard has been used also in the Latvian LinkedSAEIMA project [2], in the Italian Parliament[3] and in the PoliMedia project, where RDF parliamentary data was linked with media sources [11].

There are several parliamentary corpora. The best known is perhaps the EuroParl corpus, which includes the plenary session debates of the European Parliament and has been used to study machine translation [12]. A comprehensive list of the national parliamentary corpora is presented on the CLARIN webpage[4]. The Talk of Norway (1998–2016) is an example of a national parliament corpus with linguistic annotation published in CSV and TSV formats. [15] Different guidelines have been followed for annotating and encoding the Parliamentary debates. The TEI-based Parla-CLARIN schema, which we also use in our transformation, is an attempt to define a common annotation model.[5] For example, the Slovene parliamentary corpus siParl (1990–2018) has been encoded with the Parla-CLARIN schema [19]. Currently, the Parla-CLARIN schema is implemented in the Clarin ParlaMint project[6], which establishes a comparable and interoperable corpus of almost twenty national parliamentary corpora for comparative research.

A novelty in the transformation done in our SEMPARL project is to combine RDF standard with Parla-CLARIN schema. Moreover, most of the annotated parliamentary corpora cover mainly the recent years while in our case the complete work of the PoF from 1907 is covered—and for the first time.

3 Original Data

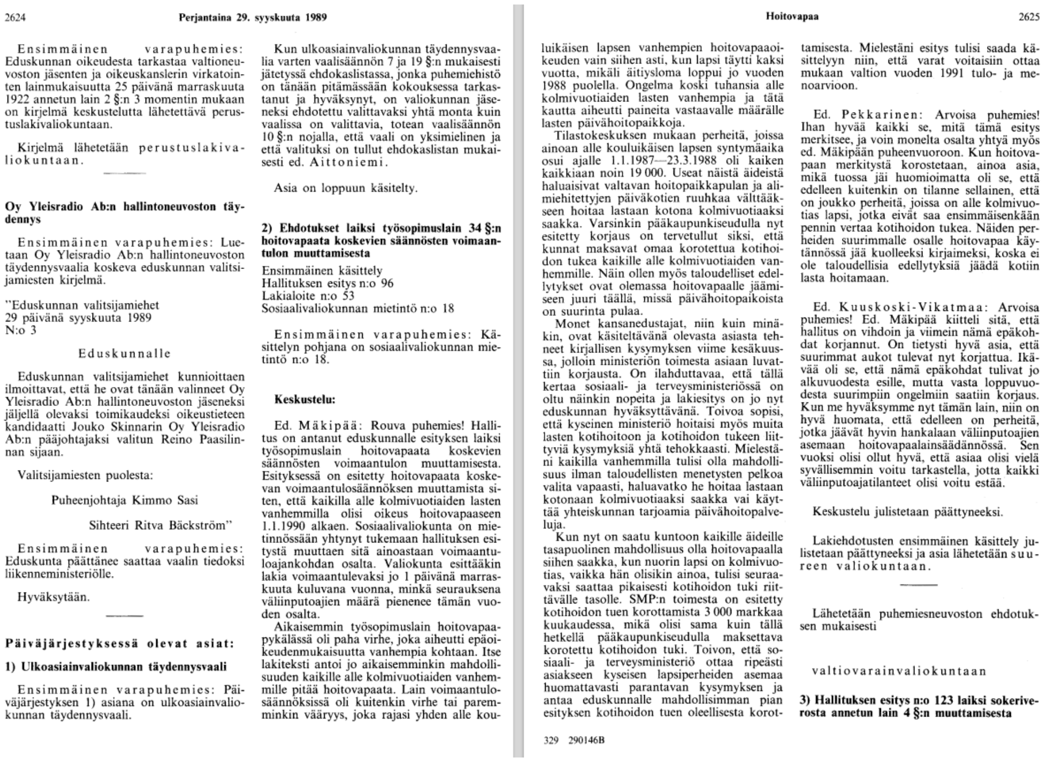
The original data, minutes of Finnish plenary sessions, was gathered from several sources and in three different formats depending on the availability: 1) From 1907 to 1999[7] the plenary

[2] https://www.w3.org/RDF/
[3] http://data.camera.it/data/en/datasets/
[4] https://www.clarin.eu/resource-families/parliamentary-corpora
[5] See: https://www.clarin.eu/blog/clarin-parlaformat-workshop
[6] https://github.com/clarin-eric/ParlaMint
[7] There is no data for 1915 and 1916 as due to war the Parliament did not convene.

session minutes are available only in PDF format[8]. One parliamentary session is split into 1–8 separate PDF files, each containing the minutes for several plenary sessions. 2) From halfway parliamentary session 1999 to the end of session 2014, the data is available also in HTML format at PoF's web pages[9]. 3) From session 2015 onward the plenary sessions are available as XML from the *Avoin eduskunta* API[10].

Figure 1 shows an example of original PDF-format minutes for plenary session 87/1989[11]. Later minutes available in HTML and XML also mostly follow shown layout and logic; In general, the minutes consist of items (or topics), marked here in bold (except the row *Keskustelu:*). The item header is followed by: a possible list of related documents, chairman's opening comments, a possible debate section marked by *Keskustelu:* (*debate/conversation*) and finally a decision and a closing statement.

Figure 1 Example of a plenary session transcript. Available by the CC BY 4.0 licence.



2624 **Perjantaina 29. syyskuuta 1989**

Ensimmäinen varapuhemies: Eduskunnan oikeudesta tarkastaa valtioneuvoston jäsenten ja oikeuskanslerin virkatointen lainmukaisuutta 25 päivänä marraskuuta 1922 annetun lain 2 §:n 3 momentin mukaan on kirjelmä keskustelutta lähetettävä perustuslakivaliokuntaan.

Kirjelmä lähetetään perustuslakivaliokuntaan.

Oy Yleisradio Ab:n hallintoneuvoston täydennys

Ensimmäinen varapuhemies: Luetaan Oy Yleisradio Ab:n hallintoneuvoston täydennysvaalia koskeva eduskunnan valitsijamiesten kirjelmä.

"Eduskunnan valitsijamiehet
29 päivänä syyskuuta 1989
N:o 3

Eduskunnalle

Eduskunnan valitsijamiehet kunnioittaen ilmoittavat, että he ovat tänään valinneet Oy Yleisradio Ab:n hallintoneuvoston jäseneksi jäljellä olevaksi toimikaudeksi oikeustieteen kandidaatti Jouko Skinnarin Oy Yleisradio Ab:n pääjohtajaksi valitun Reino Paasilinnan sijaan.

Valitsijamiesten puolesta:

Puheenjohtaja Kimmo Sasi

Sihteeri Ritva Bäckström"

Ensimmäinen varapuhemies: Eduskunta päättänee saattaa vaalin tiedoksi liikenneministeriölle.

Hyväksytään.

Päiväjärjestyksessä olevat asiat:

1) Ulkoasiainvaliokunnan täydennysvaali

Ensimmäinen varapuhemies: Päiväjärjestyksen 1) asiana on ulkoasiainvaliokunnan täydennysvaali.

Kun ulkoasiainvaliokunnan täydennysvaalia varten vaalisäännön 7 ja 19 §:n mukaisesti jätetyssä ehdokaslistassa, jonka puhemiehistö on tänään pitämässään kokouksessa tarkastanut ja hyväksynyt, on valiokunnan jäseneksi ehdotettu valittavaksi yhtä monta kuin vaalissa on valittavia, totean vaalisäännön 10 §:n nojalla, että vaali on yksimielinen ja että valituksi on tullut ehdokaslistan mukaisesti ed. Aittoniemi.

Asia on loppuun käsitelty.

2) Ehdotukset laiksi työsopimuslain 34 §:n hoitovapaata koskevien säännösten voimaantulon muuttamisesta

Ensimmäinen käsittely
Hallituksen esitys n:o 96
Lakialoite n:o 53
Sosiaalivaliokunnan mietintö n:o 18

Ensimmäinen varapuhemies: Käsittelyn pohjana on sosiaalivaliokunnan mietintö n:o 18.

Keskustelu:

Ed. Mäkipää: Rouva puhemies! Hallitus on antanut eduskunnalle esityksen laiksi työsopimuslain hoitovapaata koskevien säännösten voimaantulon muuttamisesta. Esityksessä on esitetty hoitovapaata koskevan voimaantulosäännöksen muuttamista siten, että kaikilla alle kolmivuotiaiden lasten vanhemmilla olisi oikeus hoitovapaaseen 1.1.1990 alkaen. Sosiaalivaliokunta on mietinnössään yhtynyt tukemaan hallituksen esitystä muuttaen sitä ainoastaan voimaantuloajankohdan osalta. Valiokunta esittääkin lakia voimaantulevaksi jo 1 päivänä marraskuuta kuluvana vuonna, minkä seurauksena väliinputoajien määrä pienenee tämän vuoden osalta.

Aikaisemmin työsopimuslain hoitovapaapykälässä oli paha virhe, joka aiheutti epäoikeudenmukaisuutta vanhempia kohtaan. Itse lakiteksti antoi jo aikaisemminkin mahdollisuuden kaikille alle kolmivuotiaiden vanhemmille pitää hoitovapaata. Lain voimaantulosäännöksissä oli kuitenkin virhe tai paremminkin vääryys, joka rajasi yhden alle kouluikäisen lapsen vanhempien hoitovapaaoikeuden vain siihen asti, kun lapsi täytti kaksi vuotta, mikäli äitiysloma loppui jo vuoden 1988 puolella. Ongelma koski tuhansia alle kolmivuotiaiden lasten vanhempia ja tätä kautta aiheutti paineita vastaavalle määrälle lasten päivähoitopaikkoja.

Tilastokeskuksen mukaan perheitä, joissa ainoan alle kouluikäisen lapsen syntymäaika osui ajalle 1.1.1987—23.3.1988 oli kaiken kaikkiaan noin 19 000. Useat näistä äideistä haluaisivat valtavan hoitopaikkapulan ja alimiehitettyjen päiväkotien ruuhkaa välttääkseen hoitaa lastaan kotona kolmivuotiaaksi saakka. Varsinkin pääkaupunkiseudulla nyt esitetty korjaus on tervetullut siksi, että kunnat maksavat omaa korotettua kotihoidon tukea kaikille alle kolmivuotiaiden vanhemmille. Näin ollen myös taloudelliset edellytykset ovat olemassa hoitovapaalle jäämiseen juuri täällä, missä päivähoitopaikoista on suurinta pulaa.

Monet kansanedustajat, niin kuin minäkin, ovat käsiteltävänä olevasta asiasta tehneet kirjallisen kysymyksen viime kesäkuussa, jolloin ministeriön toimesta asiaan luvattiin korjausta. On ilahduttavaa, että tällä kertaa sosiaali- ja terveysministeriössä on oltu näinkin nopeita ja lakiesitys on jo nyt eduskunnan hyväksyttävänä. Toivoa sopisi, että kyseinen ministeriö hoitaisi myös muita lasten kotihoitoon ja kotihoidon tukeen liittyviä kysymyksiä yhtä tehokkaasti. Mielestäni kaikilla vanhemmilla tulisi olla mahdollisuus ilman taloudellisten menetysten pelkoa valita vapaasti, haluavatko he hoitaa lastaan kotonaan kolmivuotiaaksi saakka vai käyttää yhteiskunnan tarjoamia päivähoitopalveluja.

Kun nyt on saatu kuntoon kaikille äideille tasapuolinen mahdollisuus olla hoitovapaalla siihen saakka, kun nuorin lapsi on kolmivuotias, vaikka hän olisikin ainoa, tulisi seuraavaksi saattaa pikaisesti kotihoidon tuki riittävälle tasolle. SMP:n toimesta on esitetty kotihoidon tuen korottamista 3 000 markkaa kuukaudessa, mikä olisi sama kuin tällä hetkellä pääkaupunkiseudulla maksettava korotettu kotihoidon tuki. Toivon, että sosiaali- ja terveysministeriö ottaa ripeästi asiakseen kyseisen lapsiperheiden asemaa huomattavasti parantavan kysymyksen ja antaa eduskunnalle mahdollisimman pian esityksen kotihoidon tuen oleellisesta korot-

329 290146B

Hoitovapaa 2625

tamisesta. Mielestäni esitys tulisi saada käsittelyyn niin, että varat voitaisiin ottaa mukaan valtion vuoden 1991 tulo- ja menoarvioon.

Ed. Pekkarinen: Arvoisa puhemies! Ihan hyvää kaikki se, mitä tämä esitys merkitsee, ja voin monelta osalta yhtyä myös ed. Mäkipään puheenvuoroon. Kun hoitovapaan merkitystä korostetaan, ainoa asia, mikä tuossa jäi huomioimatta oli se, että edelleen kuitenkin on tilanne sellainen, että on joukko perheitä, joissa on alle kolmivuotias lapsi, jotka eivät saa ensimmäisenkään pennin vertaa kotihoidon tukea. Näiden perheiden suurimmalle osalle hoitovapaa käytännössä jää kuolleeksi kirjaimeksi, koska ei ole taloudellisia edellytyksiä jäädä kotiin lasta hoitamaan.

Ed. Kuuskoski-Vikatmaa: Arvoisa puhemies! Ed. Mäkipää kiitteli sitä, että hallitus on vihdoin ja viimein nämä epäkohdat korjannut. On tietysti hyvä asia, että suurimmat aukot tulevat nyt korjattua. Ikävää oli se, että nämä epäkohdat tulivat jo alkuvuodesta esille, mutta vasta loppuvuodesta suurimpiin ongelmiin saatiin korjaus. Kun me hyväksymme nyt tämän lain, niin on hyvä huomata, että edelleen on perheitä, jotka jäävät hyvin hankalaan väliinputoajien asemaan hoitovapaalainsäädännössä. Sen vuoksi olisi ollut hyvä, että asiaa olisi vielä syvällisemmin voitu tarkastella, jotta kaikki väliinputoajatilanteet olisi voitu estää.

Keskustelu julistetaan päättyneeksi.

Lakiehdotusten ensimmäinen käsittely julistetaan päättyneeksi ja asia lähetetään suureen valiokuntaan.

Lähetetään puhemiesneuvoston ehdotuksen mukaisesti

valtiovarainvaliokuntaan

3) Hallituksen esitys n:o 123 laiksi sokeriverosta annetun lain 4 §:n muuttamisesta

Each source format differs in the metadata included. All formats contained the essential data, such as plenary session id, date, debate topic, speaker's last name, and role. The newer machine readable formats have been enriched with additional data, such as URLs to documents related to the debate topics or even individual starting and ending times for a speech. Table 1 illustrates the metadata present in each format and distribution of used

[8] https://avoindata.eduskunta.fi/#/fi/digitoidut/download
[9] https://www.eduskunta.fi/FI/taysistunto/Sivut/Taysistuntojen-poytakirjat.aspx
[10] https://avoindata.eduskunta.fi/#/fi/home
[11] https://s3-eu-west-1.amazonaws.com/eduskunta-asiakirja-original-documents-prod/suomi/1989/PTK_1989_3.pdf

source formats.

Table 1 Distribution of used source data format and variant metadata present in it. Row *Ubiquitous metadata* lists metadata that was available in all formats. * HTML became available after plenary session 85/1999.

	Parliamentary session	Speaker first name	Speaker party	Item transcript URL	Related document URL	Session transcript status	Session transcript version	Speech transcript status	Speech transcript version	Speech start time	Speech end time
PDF	1907-1999*	-	-	-	-	-	-	-	-	-	-
HTML	1999*-2014	X	X	X	X	-	-	-	-	-	-
XML	2015-2020	X	X	-	X	X	X	X	X	X	X
Ubiquitous metadata		session date, session ending and starting times, session id, speaker last name, speaker title, speech type, related documents, debate topic									

4 Target Data Model

The goal of the whole data transformation process was to make all data available in a coherent, unified format. In this project we did this twice-fold in Parla-CLARIN XML and RDF. The central unit of the data is a speech; any comment, statement or vocal contribution made during a plenary session[12]. The goal of the transformation process was to find all such speeches and all available metadata related to them. Generally we refer to all beforementioned instances as speeches. For full coverage we have also gathered all speeches made by the chairmen. These are mostly about guiding the progression of a session.

The Parla-CLARIN XML format[13] for representing speech texts is an easily readable chronological presentation of the debate data for both machines and humans. We produced one file per parliamentary session. Listing 1 gives an example of a section from the final data in Parla-CLARIN XML. The excerpt covers the start of the debate on a topic during the plenary session 37/2005.

By transforming all data to RDF as well, we aimed to create the knowledge graph (S-KG) of all parliamentary debate speeches. For this purpose a customised RDF-based metadata schema was created. The schema contains six different, interlinked classes: Speech, Interruption, Item, Session, Document, and Transcript. Speeches were represented as instances of the class Speech with 24 properties (metadata elements) as described in Table 2. Here the default namespace is our own (*semparls*); *bioc* refers to the BioCRM schema for representing biographical data [27]; *rdfs* refers to the RDFS Schema and *xsd* to the XML Schema of W3C. The column C tells the cardinality of the property, Range the range, and last column the meaning of the property. Table 3 describes in the same way the

[12] These do not include interjections, other vocal interruptions or chairman comments made during a speech. In original data these have been embedded into the actual speeches. These were handled in the transformation process as *interruptions*.

[13] https://clarin-eric.github.io/parla-clarin/

remaining five classes and additionally a seventh class, NamedEntity, that was created by post-transformation language analysis.

Listing 1 An abridged excerpt from the Parla-CLARIN data

```
<TEI xml:id="ptk_37_2005">
 [...]
  <div>
  <head>
    Eduskunnan pankkivaltuuston kertomus 2014
    <listBibl>
      <head>Related documents:</head>
      <bibl>Kertomus K 14/2015 vp</bibl>
    </listBibl>
  </head>
  <div>
    <note link=[...] speechType="" type="speaker" xml:id="2015.24.102"/>
    <u ana="#secondViceChair" who="#Paula_Risikko" xml:id="2015.24.102">
    Lähetekeskustelua varten esitellään päiväjärjestyksen 4. asia.
    Puhemiesneuvosto ehdottaa, että asia lähetetään talousvaliokuntaan
    Meille asian esittelee edustaja Zyskowicz, olkaa hyvä.</u>
  </div>
  <div>
    <note end="2015-06-24T17:54:02" link=[...] speechType="Esittelypuheenvuoro"
    start="2015-06-24T17:45:01" type="speaker" xml:id="2015.24.103"/>
    <u who="#Ben_Zyskowicz" xml:id="2015.24.103">Arvoisa rouva puhemies!
    Arvoisat kansanedustajat! Käsittelyssä on nyt Pankkivaltuuston kertomus
    vuodelta 2014. Kuten viime vuonnakin, [...] Loppuosa eli noin 137,5
    iljoonaa euroa siirrettiin valtion loputtomiin tarpeisiin.</u>
  </div>
  <div>
    <note end="2015-06-24T18:02:27" link=[...] speechType=""
    start="2015-06-24T17:54:07" type="speaker" xml:id="2015.24.104"/>
    <u next="2015.24.104.2" who="#Olavi_Ala-Nissilä" xml:id="2015.24.104.1">Arvoisa
    rouva puhemies! Tässä entinen Pankkivaltuuston puheenjohtaja, nykyinen jäsen,
    edustaja Zyskowicz käytti hyvän puheenvuoron. [...]
    Muistan, kun silloin valtiovarainministeri ja ministeri Wideroos </u>
    <vocal who="Eero_Heinäluoma">
        <desc>Eero Heinäluoma: Toinen valtiovarainministeri!</desc>
    </vocal>
    <u prev="2015.24.104.1" who="#Olavi_Ala-Nissilä" xml:id="2015.24.104.2">
    - toinen valtiovarainministeri Wideroos - ja hallituskin ajoivat sitä,
    että Suomen Pankin pääomia [...] ja Kreikan on omankin taloutensa
    kannalta välttämättä saatava julkinen hallintonsa paremmin toimimaan.</u>
    <vocal>
        <desc>Eduskunnasta: Hyvä puheenvuoro!</desc>
    </vocal>
  </div>
 [...]
```

The data model presented for representing debates is part of a larger Ontology of Parliament of Finland under development in the SEMPARL project. This ontology is based on the CIDOC CRM[14] -based Bio CRM model [27], where parliamentary events are represented in time and place with actors (people, groups, such as parties, and organizations) participating in different roles. The ontology is populated with data extracted from the speech data and databases of PoF [16]. For example, the *:speaker* and *:party* property values in Table 2 are filled with resources taken from the actor graph in the PoF ontology that contains over 2600 MPs, ministers, presidents of Finland, and other prominent people related to the speeches as speakers or mentioned in the texts. In this way, prosopographical data and the speeches can be integrated seamlessly and be used together with the Digital Humanities analyses of the parliamentary data. For example, by using biographical information about the speaker it is possible to investigate how much (s)he has spoken about matters related to his/her own electoral district.

5 Transformation Process

Semantic Parliament aggregates data from several disparate source databases into a unified knowledge graph. An overall plan of the data transformation processes of source datasets and the linking of entities between different parts are shown in Figure 2. The source datasets are shown as rectangles on the left side of the transformation pipeline and the RDF-format

[14] https://cidoc-crm.org

Table 2 Semparls RDF schema for Speech. [a]From some source data the chairmen names were not always reliably recognizable. In this case chairman speeches lack this value.

Speech			
Element URI	**C**	**Range**	**Meaning of the value**
:skos:prefLabel	1	rdf:langString	String label for speech
:speaker	0..1[a]	bioc:Person	Person speaking URI
:party	0..1	:Party	Party of the speaker URI
:partyInSource	0..1	rdfs:Literal	Party as written in the source if available
:role	0..1	:Role	Speaker's role
:speakerInSource	1	rdfs:Literal	Speaker's name as in source
:speechOrder	1	xsd:integer	Ordinal of the speech in a session
:content	1	rdfs:Literal	Speech as text (incl. interruptions)
dct:language	0..*	rdfs:Resource	Recognized languages of the speech
:speechType	0..1	:SpeechType	Type of the speech
:isInterruptedBy	0..*	:Interruption	Interruptions during the speech
dct:date	1	xsd:date	Date of the session
:startTime	0..1	xsd:time	Start time of the speech
:endDate	0..1	xsd:date	Session end date if not same as date
:endTime	0..1	xsd:time	End time of the speech
:item	0..1	:Item	Item in agenda/topic of the speech
:session	1	:Session	Session where the speech was made
:diary	1	rdfs:Resource	URL of session transcript
:page	0..1	xsd:integer	Page number for PDF-based data
:status	0..1	:Status	Status of the speech transcription
:version	0..1	xsd:decimal	Version of the speech transcription
:namedEntity	0..*	:NamedEntity	Referenced named entities
dct:subject	0..*	skos:Concept	Subject matter keywords

parts are shown as yellow cylinders. The solid arrows depict data transformation and dotted arrows correspond to entity linking either inside the Semantic Parliament data or to external ontologies and datasets (shown on the top).

The external ontologies and data shown in Figure 2 are the AMMO ontology of Finnish historical occupations, which is linked to social statuses through the international HISCO standard [13], Wikidata, related Finnish Sampo data services and portals[15], such as LawSampo [7] and BiographySampo [6], places, Finto[16] ontologies, EKS subject headings[17] used in the library of PoF, Semantic Finlex [18] data service of Finnish legislation and case law [18], and the Lakitutka[18] service publishing data related to government proposals discussed in the speeches and other documents. These will enrich the content and enhance the usefulness of the speech data for parliamentary research and applications.

The step 1 of transforming MP data is discussed in [16]. The step 2 concerning government proposals remains a future work. This paper focuses on the 3. step of the transformation of the plenary session documents and the full-text contents of the speeches given in the sessions. The entity linking from the plenary sessions to entities of the MP data is already implemented, as well as linking to places, Finto ontologies and Semantic Finlex, while linking to government proposal documents, EKS, and Lakitutka will be implemented in the future.

OCR Process In the 3. step, the data from 1907 until 1999 was available only as scanned images combined into PDF files, which needed to be first processed into machine-

[15] https://seco.cs.aalto.fi/events/2020/2020-10-29-sampo-portals/
[16] https://finto.fi/en/
[17] https://www.eduskunta.fi/kirjasto/EKS/index.html?kieli=en
[18] https://lakitutka.fi

Table 3 Semparls RDF schema for the classes Interruption, Item, Document, Session, Transcript, and ReferencedNamedEntity. Each class also contains the predicate *skos:prefLabel* that has been omitted from the table for redundancy.

Element URI	C	Range	Meaning of the value
Interruption			
:content	1	rdfs:Literal	Content of the interruption
:interrupter	0..1	rdfs:Literal	Source of the interruption
:speaker	0..1	bioc:Person	Interrupter URI, if interrupter was mentioned
Item			
:session	1	:Session	Session where item on agenda
dct:title	1	rdf:langString	Title as written in source
:relatedDocument	0..*	:Document	Document related to item
:diary	1	rdfs:Resource	URL to online transcript
Document			
dct:title	1	xsd:string	Name of the document
:id	0..1	xsd:string	Official Parliament id
:url	0..1	rdfs:Resource	URL to online transcript
Session			
:id	1	rdfs:Literal	Session id/ session number
dct:date	1	xsd:date	Date of the session
:startTime	0..1	xsd:time	Start time of the session
:endDate	0..1	xsd:date	Session end date if not same as session date
:endTime	0..1	xsd:time	End time of the session
:transcript	1	:Transcript	Transcript of the session
Transcript			
:status	0..1	:Status	Status of the transcript
:version	0..1	xsd:decimal	Version of the transcript
:url	1	rdfs:Resource	URL to online transcript
NamedEntity			
:surfaceForm	1	xsd:string	original surface forms in text
:count	1	xsd:integer	how many times entity is mentioned in a speech
:category	1	xsd:string	type of the named entity
:surfaceForm	1	xsd:string	named entity in surface form
skos:relatedMatch	0..*	rdfs:Resource	links to ontologies for named entities

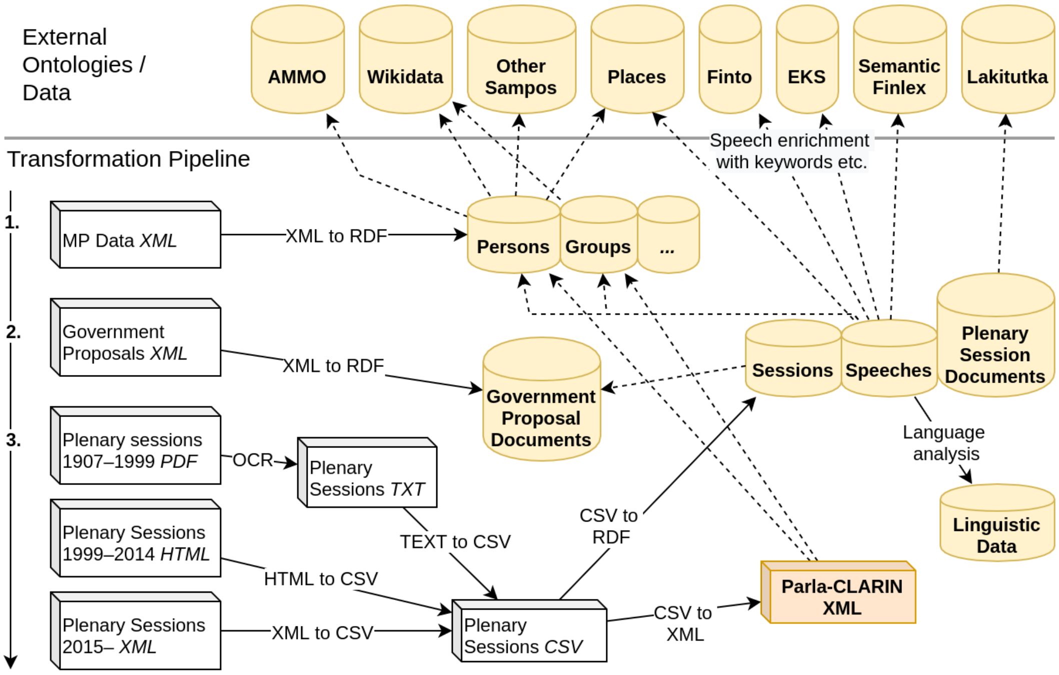
readable text. The quality of the scanned documents is generally good, with older documents having partially smudged parts of the text and some pages slightly skewed. The text in the documents is formatted into two columns, with older issues separated with a black line. There is a difference in the fonts used in different years. However, both early and later years are printed with modern fonts that are easy to recognize. Most of the text is written in Finnish, however, there are some parts written in Swedish (another official language of Finland), so we needed to use a multilingual OCR model for recognition.

For the OCR, we used Tesseract 5[19], with the default Finnish and Swedish models together for recognition `fin+swe`. The initial experiments showed that Tesseract's pre-trained models worked well with our data so we didn't need to create any training data and train new models, which simplified the whole process. Also, Tesseract's possibility to use multi-model recognition was very convenient for our dataset. As the output from the OCR process, we opted for the plain text as it seemed to be more convenient for further processing.

Since the scanned images are available in PDF files, to OCR them we needed to first transform them to PNG format. We performed the transformation with `pdftopng` program with 350 dpi resolution. In the initial experiments, we tried the OCR process with different

[19] https://github.com/tesseract-ocr, version: 5.0.0-alpha-648-gcdebe

Figure 2 Transformation process and source datasets of Semantic Parliament.



resolutions, but the 350 dpi seemed to give the best results with pre-trained OCR models.

The quality of the OCR seems to be generally good enough for our purpose. We have noticed that there are lots of mistakes in tables and lists due to Tesseract's segmentation problems. But, since we are focusing only on extracting parliamentary discussions, which are contained in the running text, we are satisfied with the OCR quality. However, during the processing of the data, we did perform some post-correction, like removing extra characters and end-of-line hyphenation, and correction of speaker names and headers.

Gathering and editing the data For the OCR-based data we decided to add one manual step to the process. Every plenary session's original minutes start with a clearly structured header row containing central information about the session (i.e. session number and date). Where the rest of the document was in most cases laid out into two columns, this header spanned both columns and was hence occasionally split or otherwise corrupted in the OCR process. To considerably improve the reliability of this central metadata, we chose to go through the files with the help of a printer script to spot these mangled headers and manually fix them. After that all relevant data was gathered with the use of regular expressions.

For the HTML-based data (step 3 in Fig. 2), we needed two steps to gather all the data. The HTML-based minutes were separated into a) a main page, listing the agenda, and links to possible debate pages and related documents, and b) possible debate pages that contained the actual debate related to an item on the agenda. Gathering the data required first scraping the main pages and then, based on the discussion page links found, the discussion data. Finally data from these sources needed to be reordered and combined into an integrated whole.

The XML-based data (2015–) was gathered with requests to *Avoin eduskunta* API that returned the minutes as JSON-wrapped XML data. The HTML- and XML-based data consisted of pre-processed elements and was mostly quite ready to use as it is. For HTML some elements did require a few string operations to split information for separate values. Regardless of the original format, all data was first transformed into CSV format, one parliamentary session a file and one speech per row with columns representing the properties of the speeches. A unique ID was created for every speech in the process.

During the history of PoF there have been cases where two parliamentary sessions refer to the same calendar year. This is due to the government resigning in the middle of a parliamentary session and hence ending the session prematurely. For example, there was the first parliamentary session in 1975 and the second parliamentary session 1975 as well. Speech and plenary session IDs related to a second parliamentary session have a *_II* suffix attached. From the year 1917 we also transformed two unofficial but historically significant meetings that took place between parliamentary sessions. These speeches, sessions, and the files containing them are marked with a *_XX* suffix.

During editing and post-correction the speeches were cleaned of original end-of-line hyphenation and other unwanted characters but the original paragraph structure was kept. The clean-up results are not yet fully perfect but already usable. Some problems, like the occasional page header texts (that have carried over from the PDF based data) remain embedded in the speech content. Post-correction was also needed for two other notable issues that, however, only concerned the PDF-based data: 1) There are cases where the speeches had been wrongly split into two with the last section having incorrect metadata. 2) Speakers who had not been recognised in the data enrichment step (to be described below in more detail) are lacking in the metadata. This was either due to the speaker's name having been corrupted in some way during the process or (more rarely) due to that the person or certain form of their name is missing from the enrichment data source or original source deviating from typical transcript convention. The aim of post-correction was to automatically spot and fix such cases.

Data enrichment During the transformations into CSV the data went through many post-corrections but also data enrichments. Most notably information about the speaker was expanded using the PoF Ontology. Where not already available in the original source, we fetched from the ontology the speaker's first name and party. If not already available in source material, we also automatically created URLs for relevant documents, such as original transcripts and related documents (bills, committee reports, etc.) if such existed. Language of each speech was checked with the LAS[20] tool.

In order to analyze the speeches and to be able to study them in more detail, the named entities in the speeches were extracted and linked to the PoF Ontology (property :referencedNamedEntity in Table 2). In order to identify named entities from the speeches, the data had to be modeled to preserve structure and interjections within the texts. The speeches were transformed into RDF, using the NIF format[21] for interoperability, separating paragraphs and titles. The interjections were identified and marked as paragraphs, so that they could be extracted from the speeches themselves. After the separation process, the data can be used for morphological analysis on the speeches and interjections separately to enable text analysis. This, however, remains as a future work.

After the speeches were transformed into RDF to preserve their structure and to separate the speeches from interjections, the RDF was used to identify named entities from the texts. The named entity extraction was done using the upgraded Nelli tool [25] and linked separately

[20] `http://demo.seco.tkk.fi/las/`
[21] `https://persistence.uni-leipzig.org/nlp2rdf/`

to be able to take the context into account. The named entities (e.g., people, places, groups and organizations) were linked internally using the ARPA tool [17], in addition to resources in external knowledge bases, such as the Kanto[22] vocabulary for Finnish actors provided by the National Library for organizations and groups, the General Finnish Ontology (YSO) for places[23] [23], PNR[24] gazetteer data of Finnish place names by the National Survey, and the Semantic Finlex[25] [18] data of the Ministry of Justice to have broader coverage for linking places, actors, and legal documents.

The subject matter keywords for each speech were extracted using Annif [24], a subject indexing tool developed by the National Library of Finland (property *dct:subject* in Table 2). The Finto REST API[26] offers Annif models that are pre-trained on categorical metadata from Finnish libraries, museums, and archives available at the Finna service[27]. These projects provide subject keywords automatically linked to entities of the General Finnish Ontology YSO. The model used for subject indexing was `yso-fi`, which combines lexical and associative approaches, so that it is able to find terms directly present in the texts as well as indirect concepts based on statistical machine learning. A list of keywords for each speech was obtained using a limit of 100 keywords and a weight threshold of 0.01.

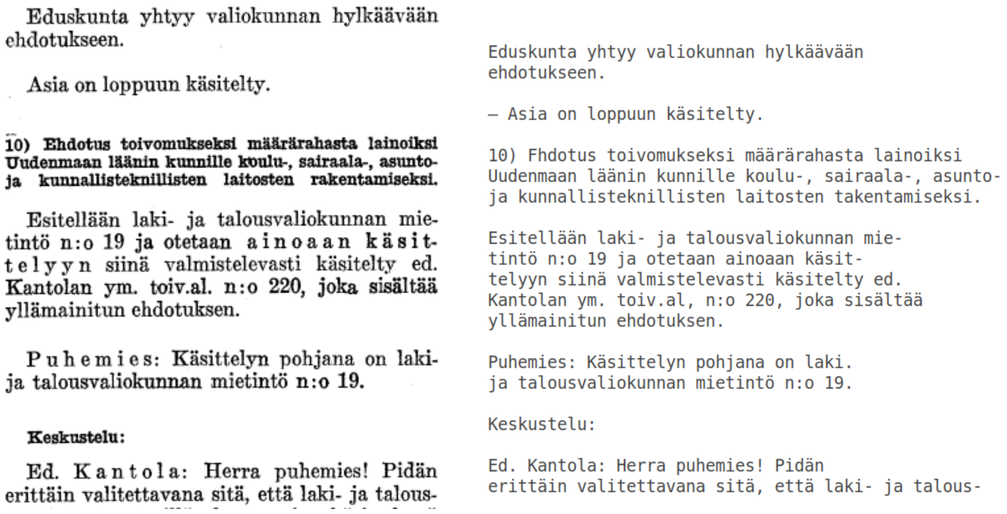
Parla-CLARIN Transformation The transformation to Parla-CLARIN was a fairly straight-forward process of creating an XML tree from the CSV data. Each file, containing one parliamentary session, forms its own entity, containing all session and speaker metadata with proper ID-linkage inside the document. We chose to separate all interruptions from the actual speech content by separating them to their own elements (as seen in Listing 1).

RDF Transformation From the initial CSV, the debates were also transformed into RDF. For this we used the Terse RDF Triple Language (Turtle) syntax[28] and the schema presented in Section 4. The data for one parliamentary session was recreated as three different interlinked files, the first containing all the actual speeches made during that whole parliamentary session and all immediate metadata such as information about the speaker and the date. These link to a second file containing all the items discussed and related documents and their available metadata. The third file consists of the parliamentary session's plenary sessions and minutes transcripts. In the forming of URIs for the people and parties we once again utilized the PoF Ontology to ensure fluent linkage between the speech and prosopographical data sets.

6 Validation

The whole process extracted over 900 000 individual speeches from the whole period, from 1907 to current day. The length of a speech can vary from a single word to over thousand words in length. A completely automated process handling this much data is naturally prone to errors in dealing with exceptions in the data. At this point most validation of the result data has been manual. Currently, we are looking more deeply into the OCR results to get more concrete understanding of our success in that step of the process. Fig. 3 shows a snippet of the data in the original PDF format used and in the final text form. Apart from issues

[22] https://finto.fi/finaf/en/
[23] https://finto.fi/yso-paikat/en/?clang=en
[24] http://www.ldf.fi/dataset/pnr
[25] https://data.finlex.fi
[26] http://api.finto.fi/
[27] http://www.finna.fi
[28] https://www.w3.org/TR/turtle/



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-
```

(a) Source PDF transcript of plenary session 49/1967, p. 885 **(b)** Result text after OCR

Figure 3 Example of the source and the result after the OCR process.

described in section 5, Transformation Process, we have observed that the quality of the OCR results vary from decade to decade. The quality of 1990's OCR is quite good, with very little issues on relevant parts while results from the start of the 20^{th} century contain more errors. The main reason for these differences is the varying quality of available images and the paper the original document was printed on. A similar trend has been observed in [14].

Preliminary tests on speaker recognition (i.e., that each speech has speaker property value with speaker name and other required speaker metadata associated with it) show that after corrections the amount of recognized speakers tends to be over 99%. These tests were performed on random parliamentary sessions from all OCR-based decades. It is good to note that these numbers do not indicate whether the speaker is the correct one, as in some cases the chance of incorrect name correction or split speech does remain.

The RDF data model of the parliamentary debates is presented in a machine-processable format using the ShEx Shape Expressions language[29] [26]. We have made initial validation experiments with PyShEx[30] and shex.js[31] validators. Based on the experiments, we have identified errors both in the schema and the data. The schema errors include syntax errors, incorrect cardinality definitions, incorrect literal datatype definitions, and incorrect namespaces for IRI values. The errors in the schema have been fixed accordingly. In the data, we have found systematic issues stemming from the RDF conversion process, e.g., some separate speeches and interruptions that were merged into one speech/interruption instance, speeches that were attached to multiple session item and diary (should be only one), and triples with an incorrectly minted object IRI (the base IRI of the Turtle file) instead of omitting the value altogether. The issues have been fixed in the data conversion process. We plan a full-scale ShEx validation phase integrated in the data conversion and publication process to spot and report errors in the dataset.

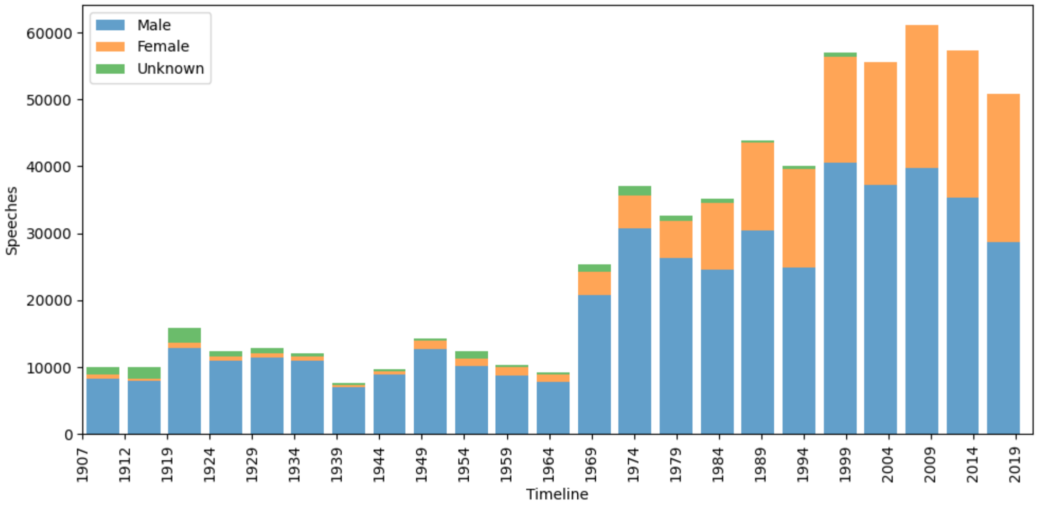
[29] https://shex.io
[30] https://github.com/hsolbrig/PyShEx
[31] https://github.com/shexSpec/shex.js

7 Publishing and Using Speeches via a Linked Open Data Service

The S-KG has been published on the Linked Data Finland platform[32] [8] according to the Linked Data publishing principles and other best practices of W3C [4], including, e.g., content negotiation and provision of a SPARQL[33] endpoint[34].

The data will be used via the SPARQL endpoint in two ways. Firstly, a portal called *ParliamentSampo – Finnish Parliament on the Semantic Web* is under development, a new member in the Sampo series of semantic portals[35]. The portal includes data analytic tools studying parliamentary debates, networks of Finnish politicians, and political culture, and is targeted to both researchers and the public for. Secondly, in addition to the ready-to-use application perspectives in the ParliamentSampo portal, the underlying SPARQL endpoint can and is being applied to custom data analyses in Digital Humanities research using YASGUI[36] [22] and Python scripting in Google Colab[37] and Jupyter[38] notebooks. In our work, the "FAIR guiding principles for scientific data management and stewardship" of publishing Findable, Accessible, Interoperable, and Re-usable data are used[39].

Figure 4 Total number of speeches by gender.



One example of using the data for analysis through SPARQL endpoint is shown in Fig. 4. It represents the number of speeches on a timeline by gender. The histogram shows the speeches of male speakers with a blue bar and female speakers with an orange bar. The green bar is for speeches where the speaker has not been identified due to speaker recognition issues described earlier. The chairpersons have been filtered out as they are often mentioned

[32] https://ldf.fi
[33] https://www.w3.org/TR/sparql11-query/
[34] Access to this and the Parla-CLARIN dataset is currently restricted to consortium members.
[35] https://seco.cs.aalto.fi/applications/sampo/
[36] https://yasgui.triply.cc
[37] https://colab.research.google.com/notebooks/intro.ipynb
[38] https://jupyter.org
[39] https://www.go-fair.org/fair-principles/

by the title in the data and therefore cannot be linked based on the speaker data to the actor data. With this in mind, it can be seen from the plot that the number of female speeches rises with time.

8 Discussion & Conclusions

This paper presented the first homogeneous publication of the full set of plenary speeches of PoF (1907–present) as a knowledge graph (S-KG) as Linked Data and in the emerging Parla-CLARIN standard. Thus far the speeches have been available only in PDF form, as text, in HTML, or in XML form depending on the time period and data publication.

Unlike in many other similar projects we have not focused only on a slice of existing data. Instead we have covered and brought into an unified format the speeches from the whole of Parliament of Finland's history. This makes it possible for any research to easily cover all of history with a single query and brings about completely new possibilities for further data analysis and research.

The main technical novelties in our approach w.r.t the related works discussed in Section 2 include the combined model of Parla-CLARIN and RDF developed for representing the speeches, integration of the data to the larger PoF Ontology for deeper data analyses, and enriching the data with a variety of external related national data sources to earn the 5th star according to the Linked Data 5-star model[40].

The variety of the pre-existing source formats is a key motivator for our work but also naturally a challenge. Bringing about a harmonious dataset from different sources is not a simple matter and requires familiarity with the source data. To deepen our understanding, we have also reached out to the Parliament's Central Office staff who are responsible for creating the minutes. This co-operation has been very beneficial.

The data has been published on the Linked Data Finland platform and is being used in Digital Humanities Research for studying the parliamentary language and political culture in the SEMPARL project and for implementing the end user applications. To earn the 6th star in Linked Data Finland model extending the 5-star model for better re-usability, the schema has been included and documented as part of the data publication, and to some extent validated for the 7th star. The Parla-CLARIN data set has also been already taken into internal use in the consortium and while still undergoing revision, both data sets have proved promising and fit for use. The data and data service will be used also in the Helsinki Digital Humanities Hackathon[41] in May 2021 for feedback from external users. FinnParla data will eventually be opened during the SEMPARL project by the open license CC BY 4.0.

The S-KG data will be used as a basis of the semantic portal *ParliamentSampo – Finnish Parliament on the Semantic Web* that is being developed in the Semantic Parliament project, based on the Sampo model [5] and Sampo-UI framework [9]. The Parla-CLARIN version will also be made available to the public.

Regarding data enrichment, improvements in the keyword extraction mechanism as well as automatic recognition of broad topics in the dataset are planned for the near future. We also aim to further the combination of both presented formats by creating a third version of the data as LOD using Parla-CLARIN markup for the speech contents.

Acknowledgments Thanks to Ari Apilo, Sari Wilenius, and Päivikki Karhula of PoF for providing material for the project. Our work was funded by the Academy of Finland as

[40] `https://www.w3.org/community/webize/2014/01/17/what-is-5-star-linked-data/`
[41] `http://heldig.fi/dhh21`

part of the Semantic Parliament project, the EU project InTaVia: In/Tangible European Heritage[42], and is related to the COST action NexusLinguarum[43] on linguistic data science. CSC – IT Center for Science, Finland, provided computational resources for the work.

References

1 Kaspar Beelen, Timothy Alberdingk Thijm, Christopher Cochrane, Kees Halvemaan, Graeme Hirst, Michael Kimmins, Sander Lijbrink, Maarten Marx, Nona Naderi, Ludovic Rheault, and et al. Digitization of the canadian parliamentary debates. *Canadian Journal of Political Science*, 50(3):849–864, 2017. `doi:10.1017/S0008423916001165`.

2 Uldis Bojārs, Roberts Darģis, Uldis Lavrinovičs, and Pēteris Paikens. LinkedSaeima: A linked open dataset of Latvia's parliamentary debates. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 50–56, Cham, 2019. Springer–Verlag.

3 Eduskunta. Eduskunnan täysistunnot, ladattava versio 1.5, 2017. URL: `http://urn.fi/urn:nbn:fi:lb-2019101721`.

4 Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space (1st edition)*. Morgan & Claypool, Palo Alto, California, 2011. URL: `http://linkeddatabook.com/editions/1.0/`.

5 Eero Hyvönen. "Sampo" model and semantic portals for digital humanities on the semantic web. In *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pages 373–378. CEUR Workshop Proceedings, vol. 2612, October 2020. URL: `http://ceur-ws.org/Vol-2612/poster1.pdf`.

6 Eero Hyvönen, Petri Leskinen, Minna Tamper, Heikki Rantala, Esko Ikkala, Jouni Tuominen, and Kirsi Keravuori. Biographysampo - publishing and enriching biographies on the semantic web for digital humanities research. In Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J.G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar, editors, *The Semantic Web. ESWC 2019*, pages 574–589. Springer-Verlag, June 2019. `doi:10.1007/978-3-030-21348-0\_37`.

7 Eero Hyvönen, Minna Tamper, Arttu Oksanen, Esko Ikkala, Sami Sarsa, Jouni Tuominen, and Aki Hietanen. LawSampo: A semantic portal on a linked open data service for finnish legislation and case law. In *The Semantic Web: ESWC 2020 Satellite Events. Revised Selected Papers*, pages 110–114. Springer–Verlag, 2019.

8 Eero Hyvönen, Jouni Tuominen, Miika Alonen, and Eetu Mäkelä. Linked Data Finland: A 7-star model and platform for publishing and re-using linked datasets. In *ESWC 2014 Satellite Events*, pages 226–230. Springer-Verlag, 2014.

9 Esko Ikkala, Eero Hyvönen, Heikki Rantala, and Mikko Koho. Sampo-UI: A Full Stack JavaScript Framework for Developing Semantic Portal User Interfaces. *Semantic Web – Interoperability, Usability, Applicability*, 2021. accepted.

10 Kimmo Kettunen and Matti La Mela. Digging deeper into the finnish parliamentary protocols – using a lexical semantic tagger for studying meaning change of everyman's rights (allemansrätten). In *DHN 2020 Digital Humanities in the Nordic Countries. Proceedings of the Digital Humanities in the Nordic Countries 5th Conference*, pages 63–80. CEUR Workshop Proceedings, vol. 2612, October 2020. URL: `http://ceur-ws.org/Vol-2612/paper5.pdf`.

11 Martijn Kleppe, Laura Hollink, Max Kemman, Damir Juric, Henri Beunders, Jaap Blom, Johan Oomen, and Geert-Jan Houben. Polimedia: Analysing media coverage of political debates by automatically generated links to radio & newspaper items. In *OKCon 2013 LinkedUp Veni Competition on Linked and Open Data for Education*, pages 63–80. CEUR

[42] `https://intavia.eu`

[43] `https://nexuslinguarum.eu`

Workshop Proceedings, vol. 1124, September 2013. URL: http://ceur-ws.org/Vol-1124/linkedup_veni2013_04.pdf.

12 Philipp Koehn. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86, 2005. URL: https://homepages.inf.ed.ac.uk/pkoehn/publications/europarl-mtsummit05.pdf.

13 Mikko Koho, Lia Gasbarra, Jouni Tuominen, Heikki Rantala, Ilkka Jokipii, and Eero Hyvönen. AMMO Ontology of Finnish Historical Occupations. In *Proceedings of the First International Workshop on Open Data and Ontologies for Cultural Heritage (ODOCH'19)*, volume 2375, pages 91–96. CEUR Workshop Proceedings, June 2019. URL: http://ceur-ws.org/Vol-2375/.

14 Matti La Mela. Tracing the emergence of nordic allemansrätten through digitised parliamentary sources. In Mats Fridlund, Mila Oiva, and Petri Paju, editors, *Digital histories: Emergent approaches within the new digital history*, pages 181–197. Helsinki University Press, 2020. doi:10.33134/HUP-5-11.

15 Emanuele Lapponi, Martin G. Søyland, Erik Velldal, and Stephan Oepen. The Talk of Norway: a richly annotated corpus of the Norwegian parliament, 1998–2016. *Language Resources and Evaluation*, 52(3):873–893, September 2018. doi:10.1007/s10579-018-9411-5.

16 Petri Leskinen, Jouni Tuominen, and Eero Hyvönen. Members of parliament in finland (1907–) knowledge graph and its linked open data service, 2021. Submitted for review.

17 Eetu Mäkelä. Combining a REST lexical analysis web service with SPARQL for mashup semantic annotation from text. In *Proceedings of the ESWC 2014 demonstration track*, pages 424–428. Springer-Verlag, 2014. doi:10.1007/978-3-319-11955-7_60.

18 Arttu Oksanen, Jouni Tuominen, Eetu Mäkelä, Minna Tamper, Aki Hietanen, and Eero Hyvönen. Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web. In G. Peruginelli and S. Faro, editors, *Knowledge of the Law in the Big Data Age*, volume 317 of *Frontiers in Artificial Intelligence and Applications*, pages 212–228. IOS Press, 2019.

19 Andrej Pancur and Tomaž Erjavec. The siParl corpus of Slovene parliamentary proceedings. In *Proceedings of the Second ParlaCLARIN Workshop*, pages 28–34, Marseille, France, May 2020. European Language Resources Association. URL: https://www.aclweb.org/anthology/2020.parlaclarin-1.6.

20 Onni Pekonen. *Debating "the ABCs of parliamentary life": the learning of parliamentary rules and practices in the late nineteenth-century Finnish Diet and the early Eduskunta*. PhD thesis, University of Jyväskylä, Jyväskylä, 2014. URL: http://urn.fi/URN:ISBN:978-951-39-5843-5.

21 Christian Rauh, Pieter De Wilde, and Jan Schwalbach. The ParlSpeech data set: Annotated full-text vectors of 3.9 million plenary speeches in the key legislative chambers of seven European states, 2017. doi:10.7910/DVN/E4RSP9.

22 Laurens Rietveld and Rinke Hoekstra. The YASGUI family of SPARQL clients. *Semantic Web*, 8(3):373–383, 2017.

23 Katri Seppälä and Eero Hyvönen. Asiasanaston muuttaminen ontologiaksi. Yleinen suomalainen ontologia esimerkkinä FinnONTO-hankkeen mallista (Changing a keyword thesaurus into an ontology. General Finnish Ontology as an example of the FinnONTO model). Technical report, National Library, Plans, Reports, Guides, March 2014. URL: https://www.doria.fi/handle/10024/96825.

24 Osma Suominen. Annif: DIY automated subject indexing using multiple algorithms. *LIBER Quarterly*, 29(1):1–25, 2019. URL: http://www.liberquarterly.eu/article/10.18352/lq.10285/, doi:10.18352/lq.10285.

25 Minna Tamper, Arttu Oksanen, Jouni Tuominen, Aki Hietanen, and Eero Hyvönen. Automatic annotation service APPI: Named entity linking in legal domain. In *Proceedings of ESWC 2020, Posters and Demos*. Springer–Verlag, 2020.

26 Katherine Thornton, Harold Solbrig, Gregory S. Stupp, Jose Emilio Labra Gayo, Daniel Mietchen, Eric Prud'hommeaux, and Andra Waagmeester. Using shape expressions (ShEx)

to share RDF data models and to guide curation with rigorous validation. In Pascal Hitzler, Miriam Fernández, Krzysztof Janowicz, Amrapali Zaveri, Alasdair J.G. Gray, Vanessa Lopez, Armin Haller, and Karl Hammar, editors, *The Semantic Web. ESWC 2019*, pages 606–620. Springer-Verlag, 2019. doi:10.1007/978-3-030-21348-0_39.

27 Jouni Tuominen, Eero Hyvönen, and Petri Leskinen. Bio CRM: A data model for representing biographical data for prosopographical research. In *Biographical Data in a Digital World (BD2017)*, 2017. URL: https://doi.org/10.5281/zenodo.1040712.

28 Astrid van Aggelen, Laura Hollink, Max Kemman, Martijn Kleppe, and Henri Beunders. The debates of the European Parliament as Linked Open Data. *Semantic Web*, 8(2):271–281, 2017. doi:10.3233/SW-160227.

29 Eero Voutilainen. Tekstilajitietoista kielenhuoltoa: puheen esittäminen kirjoitettuna eduskunnan täysistuntopöytäkirjoissa. In Liisa Tiittula and Pirkko Nuolijärvi, editors, *Puheesta tekstiksi – Puheen kirjallisen esittämisen alueita, keinoja ja rajoja*, pages 162–191. Suomalaisen Kirjallisuuden Seura, 2016.