




# LinkedSaeima: A Linked Open Dataset of Latvia's Parliamentary Debates

Uldis Bojārs<sup>1,2</sup> , Roberts Dargis<sup>2</sup>, Uldis Lavrinovičs<sup>3</sup>,  
and Pēteris Paikens<sup>2</sup> 

<sup>1</sup> Faculty of Computing, University of Latvia,  
Raina bulvaris 19, Riga 1459, Latvia  
uldis.bojars@lu.lv

<sup>2</sup> Institute of Mathematics and Computer Science, University of Latvia,  
Raina bulvaris 29, Riga 1459, Latvia  
{roberts.dargis, peteris.paikens}@lumii.lv

<sup>3</sup> SIA LETA Innovation Labs, Marijas iela 2, Riga 1050, Latvia  
uldis.lavrinovics@leta.lv

**Abstract.** This paper describes the LinkedSaeima dataset that contains structured data about Latvia's parliamentary debates from 1993 until 2017. This information is published at <http://dati.saeima.korpuss.lv> as Linked Open Data. It is a part of the Corpus of Saeima (the Parliament of Latvia) released as open data for multidisciplinary research. The data model of LinkedSaeima follows the data structure of the LinkedEP dataset with a few modifications. The dataset is augmented with links to the Wikidata knowledge base that provide additional information about the speakers and named entities mentioned in the corpus.

**Keywords:** Linked Open Data · Parliament debate corpus · Named entity linking · Open government data · RDF

## 1 Introduction

To ensure transparency of political and legislative processes, parliament proceedings and debate transcripts are usually made public. Saeima – the Parliament of the Republic of Latvia – publishes plenary transcripts on its website as unstructured text<sup>1</sup>. In 2016 we published this as a text corpus with speaker annotations and other metadata [1].

With the increasing availability of corpora in different languages we realized that unannotated corpora are not enough to address various researchers' needs such as comparative research across multiple languages. The 2018 release of the Corpus of Saeima attempted to address this concern by adding multiple additional annotation layers including named entity mentions, automated English translation and morphosyntactic information for linguistic analysis [2]. This release is available in multiple

<sup>1</sup> <http://www.saeima.lv/lv/transcripts/category/21>.

commonly used formats: as a text corpus in NoSketch query software<sup>2</sup>, as syntactically parsed data and as Linked Open Data [3].

This paper describes LinkedSaeima<sup>3</sup> – a Linked Data representation of the Corpus of Saeima containing structured information about Saeima proceedings and the entities mentioned in the proceedings, represented using Wikidata identifiers [4]. Linked Data allows us to represent structured information about parliamentary debates by describing the properties of the objects from the domain of parliamentary meetings and relations between these objects.

## 2 Parliamentary Speech Corpus

The source of data for this corpus is the Saeima website that contains transcripts of all parliament sessions in text format. These transcripts are processed using a semi-automatic pipeline to identify the boundaries of speeches and the speakers.

The Corpus of Saeima contains information about debates from seven parliamentary terms (5th–12th Saeima) covering years 1993–2017. The transcriptions of this corpus contain 38 million tokens and 497 thousand utterances. The available metadata for each utterance includes the date and type of the parliamentary session and speakers' names and affiliations. A subset of speeches, starting from 2015, were translated from Latvian to English using a neural machine translation system [5]. The unreviewed machine-generated translation is included in the corpus for quantitative analysis purposes and to aid searchability and understanding for international researchers. However, the text quality of automated translation is not sufficient for qualitative analysis of the Saeima corpus.

The named entities mentioned in this corpus were automatically linked to Wikidata as the entity knowledge base [4]. The named entity recognition system is based on a full text search of Wikidata entity names, extending these aliases by generating a heuristic list of alternative variants for organization and people names, and inflecting them through a custom Latvian phrase inflection system built upon the Latvian morphosyntactic tagger [6]. As the goal of named entity recognition was primarily to provide a mapping to Wikidata, no technical means were applied to recognize entities without relevant Wikidata entries, however, in order to improve the coverage of entity linking, Wikidata entries for historical members of parliament and other officials were created (if not already existing) and populated with data based on open access sources available from Saeima. For the purposes of disambiguation of entities with overlapping names, the most likely entity was chosen based on a cosine similarity metric with respect to structured Wikidata information extracts, adapting a system developed earlier for news corpora analysis [7].

---

<sup>2</sup> NoSketch interface for this corpus: <http://dati.saeima.korpuss.lv/nosketch/>.

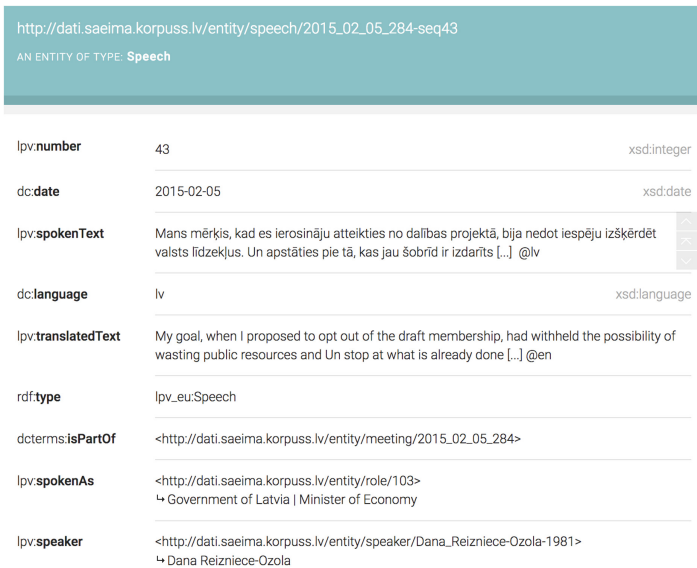
<sup>3</sup> <http://dati.saeima.korpuss.lv/>.

### 3 LinkedSaeima Dataset

This paper focuses on LinkedSaeima – the Linked Data representation of the Saeima speech corpus. The current version of the dataset, published in May 2019, consists of approx. 4.9 million RDF triples<sup>4</sup>. Since the original January 2018 release we have fixed the identified issues with its RDF representation and improved the usability of the human-readable view of the dataset.

The dataset contains 497221 speeches (utterances) from 1293 parliament meetings. These speeches were given by 690 speakers with 162 speaker roles and contain 392530 mentions of 2998 unique Wikidata entities. It includes information about the following classes of objects:

- Meeting (*lpv\_eu:SessionDay*) – a top-level concept representing one parliament plenary meeting usually consisting of multiple Speeches;
- Speech (*lpv\_eu:Speech*) – an individual speech (utterance) given at a Meeting by a single Speaker in a particular Role;
- Speaker (*lpv:Speaker*) – a person giving a speech;
- Role (*lpv:PoliticalFunction*) – a role which the person represented when giving a Speech (e.g. the Prime Minister). A person may appear in multiple roles.



http://dati.saeima.korpuss.lv/entity/speech/2015_02_05_284-seq43		
AN ENTITY OF TYPE: <b>Speech</b>		
lpv.number	43	xsd:integer
dc.date	2015-02-05	xsd:date
lpv.spokenText	Mans mērķis, kad es ierosināju atteikties no dalības projektā, bija nedot iespēju izšķērdēt valsts līdzekļus. Un apstāties pie tā, kas jau šobrīd ir izdarīts [...] @lv	
dc.language	lv	xsd:language
lpv.translatedText	My goal, when I proposed to opt out of the draft membership, had withheld the possibility of wasting public resources and Un stop at what is already done [...] @en	
rdf.type	lpv_eu:Speech	
dcterms:isPartOf	<http://dati.saeima.korpuss.lv/entity/meeting/2015_02_05_284>	
lpv.spokenAs	<http://dati.saeima.korpuss.lv/entity/role/103> ↳ Government of Latvia   Minister of Economy	
lpv.speaker	<http://dati.saeima.korpuss.lv/entity/speaker/Dana_Reizniece-Ozola-1981> ↳ Dana Reizniece-Ozola	

**Fig. 1.** LinkedSaeima information about a speech (in LodView browser). ([http://dati.saeima.korpuss.lv/entity/speech/2015\\_02\\_05\\_284-seq43](http://dati.saeima.korpuss.lv/entity/speech/2015_02_05_284-seq43))

<sup>4</sup> LinkedSaeima RDF dump is available at <http://saeima.korpuss.lv/datasets/rdf/>.

Figure 1 shows an example of a Speech. Its properties include date (*dc:date*), sequence number (*lpv:number*), spoken (*lpv:spokenText*) and translated (*lpv:translatedText*) text, and it is related to the SessionDay it is a part of (*dct:isPartOf*), to the Speaker (*lpv:speaker*), its Role (*lpv:spokenAs*) and to the named entities recognized in the text.

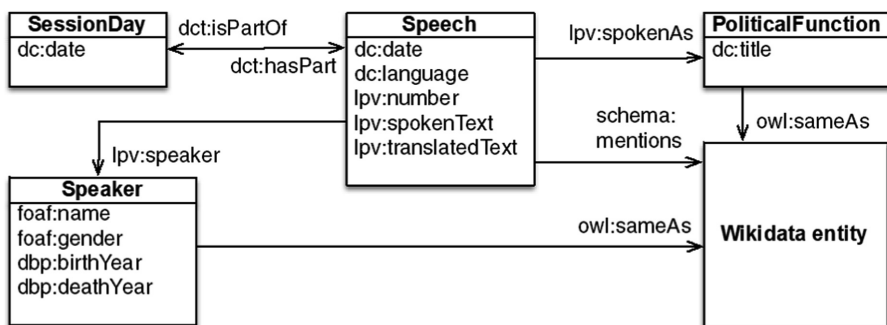


Fig. 2. The data model of the LinkedSaeima dataset.

The data model of the LinkedSaeima dataset, shown in Fig. 2, follows the model of the LinkedEP project and the Linkedpolitics vocabulary used in it, referenced in this paper using vocabulary prefixes *lpv* and *lpv\_eu* [8]. The main innovation of this dataset, compared to LinkedEP, is the addition of named entity information, represented using the *schema:mentions* property pointing to entity Wikidata identifiers. Another difference is that we “materialize” speaker Roles extracted from the corpus by giving them URI identifiers that can be used for querying the dataset (e.g. for speeches by Ministers of Foreign Affairs) and linking them to other datasets. Speaker roles (*lpv:PoliticalFunction*) may also contain links to matching entities in Wikidata.

There is ongoing work for standardization of corpora of parliamentary proceedings based on TEI [9]. Our approach could be applied to other parliamentary speech corpora by implementing a transformation from the TEI standard once it is finalized in order to make these resources available as Linked Data.

## 4 Data Access and Implementation

The LinkedSaeima dataset can be accessed:

- as Linked Data (published using LodView);
- using a Triple Pattern Fragments server and user interface<sup>5</sup>;
- as a single RDF file<sup>6</sup>.

<sup>5</sup> <http://dati.saeima.korpuss.lv/ldf/saeima>.

<sup>6</sup> <http://saeima.korpuss.lv/datasets/rdf/>.

The dataset is published as Linked Data, all its objects have HTTP URIs and information about them can be retrieved by looking up their URIs. The Linked Data interface is implemented using the LodView linked data browser<sup>7</sup> that can serve data in RDF, HTML and multiple other formats. The URI patterns used in the dataset, illustrated by examples, are listed in Table 1.

In order to provide a lightweight query interface, the dataset is published using the Triple Pattern Fragments (TPF) server which provides a lightweight way for querying RDF datasets [10]. The dataset is also released as a single RDF file that researchers can use to run more complex queries and analysis. For example, Listing 1 demonstrates how researchers can use SPARQL to perform statistical queries on this dataset.

**Table 1.** URI patterns used in the LinkedSaeima dataset.

Type	URI pattern
Speech	/entity/speech/2015_02_05_284-seq 43
Speaker	/entity/speaker/Dana_Reizniece-Ozola-1981
Role	/entity/role/103
SessionDay	/entity/meeting/2015_02_05_284

**Listing 1.** A query for the yearly statistics of speeches by the Minister of Foreign Affairs

```
PREFIX lpv: <http://purl.org/linkedpolitics/vocabulary/>
PREFIX lpv_eu: <http://purl.org/linkedpolitics/vocabulary/eu/plenary/>
PREFIX saeima_role: <http://dati.saeima.korpuss.lv/entity/role/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>

SELECT ?year (COUNT(?speech) AS ?count)
WHERE {
  ?speech a lpv_eu:Speech .
  ?speech lpv:spokenAs saeima_role:23 .
  ?speech dc:date ?date .
  BIND (year(?date) as ?year) .
}
GROUP BY ?year
ORDER BY ?year
```

<sup>7</sup> LodLive linked data browser: <https://github.com/dvcama/LodLive/>.

## 5 Conclusions

In this paper we described LinkedSaeima – a Linked Data representation of the dataset of Latvia’s parliamentary debates extended with NLP annotation layers. We hope that its Linked Data representation and the new annotation levels (entity references and translation) will allow researchers from other countries to use this resource in their studies, comparing Latvia’s parliamentary data with data from other national parliaments and to provide users with new ways of exploring this information.

Expected future work includes extending the LinkedSaeima dataset with additional types of structured information, for example, voting data, and adding automated translations for the whole historical dataset. Improvements to entity recognition and morphosyntactic tagging are being carried out as part of related research projects.

By publishing this parliamentary corpus as Linked Open Data and by including links to Wikidata entities we hope to facilitate the development of a global network of linked political and legal information, and to provide an example to other implementers.

**Acknowledgments.** This research has been partially supported by the University of Latvia project AAP2016/B032 “Innovative information technologies”, the European Regional Development Fund under the grant agreement No. 1.1.1.1/16/A/219 and the research project “Competence Centre of Information and Communication Technologies” of EU Structural funds, IT Competence Centre contract No. 1.2.1.1/18/A/003 research project No. 2.4 “Platform for the semantically structured information extraction from the massive Latvian news archive”.

## References

1. Dargis, R., Rābante-Buša, G., Auziņa, I., Kruks, S.: ParliSearch - A system for large text corpus discourse analysis. *Frontiers in Artificial Intelligence and Applications*, vol. 289, pp. 115–121 (2016)
2. Dargis, R., Auziņa, I., Bojārs, U., Paikens, P., Znotiņš, A.: Annotation of the corpus of the Saeima with multilingual standards. In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)* (2018)
3. Heath, T., Bizer, C.: *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web, 1st edn., vol. 1, no. 1, pp. 1–136. Morgan & Claypool (2011)
4. Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., Hellmann, S.: Wikidata through the Eyes of DBpedia. *Semant. Web* 9(4), 493–503 (2018)
5. Barone, A.V.M., Helcl, J., Sennrich, R., Haddow, B., Birch, A.: Deep architectures for neural machine translation. In: *Proceedings of the Second Conference on Machine Translation, Vol. 1: Research Papers*, pp. 99–107. Association for Computational Linguistics (2017)
6. Paikens, P.: Deep neural learning approaches for Latvian morphological tagging. In: *Proceedings of Human Language Technologies - The Baltic Perspective*, pp. 119–125 (2016)
7. Paikens, P.: Latvian newswire information extraction system and entity knowledge base. In: *Proceedings of Human Language Technologies - The Baltic Perspective*, pp. 119–125 (2014)

8. van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., Beunders, H.: The debates of the European parliament as linked open data. *Semant. Web* **8**(2), 271–281 (2017)
9. Erjavec, T., Pančur, A.: Parla-CLARIN: a TEI schema for corpora of parliamentary proceedings (2019). <https://clarin-eric.github.io/parla-clarin/>
10. Verborgh, R., et al.: Triple pattern fragments: a low-cost knowledge graph interface for the web. *J. Web Semant.* **37**, 184–206 (2016)

**Open Access** This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

