LinkedSaeima:

Linked Open Dataset of Latvia's Parliamentary Debates

Uldis Bojārs, Roberts Darģis
Uldis Lavrinovičs, Pēteris Paikens

NATIONAL DEVELOPMENT PLAN 2020
EUROPEAN UNION
European Regional Development Fund
INVESTING IN YOUR FUTURE

AiLab
Mākslīgā intelekta laboratorija
LU MII

LATVIJAS UNIVERSITATE
ANNO 1919
UNIVERSITY OF LATVIA

LinkedSaeima

The corpus of Latvia's parliamentary debates

- ... published as Linked Data
- ... based on the LinkedEP * model

http://dati.saeima.korpuss.lv/

Motivation: Offering new ways how to look at the parliament debates and how explore them.

* van Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., Beunders, H.: The debates of the European parliament as linked open data. Semantic Web, 8(2), 271–281 (2017)

Corpus of the Saeima

Saeima is the Parliament of the Republic of Latvia.

The Corpus of the Saeima is a collection of transcriptions of parliamentary debates from 7 parliamentary terms (5th–12th) covering years 1993 – 2017. It contains 38 million tokens and 497 thousand utterances (speeches).

Enrichment Layers

- **English translation**
- **Named entities**
- Morphological and syntactical annotations

Available datasets

- Bonito corpus browser
- Universal Dependencies format
- **LinkedSaeima (Linked Data)**

Sēdes vadītāja. Un mēs turpinām ar deputātu Sudrabas, Šimfas, Meijas, Platpera, Kūtra un Baloža jautājumu Ministru prezidentei Laimdotai Straujumai, kas ir pāradresēts ekonomikas ministrei Danai Reizniecei-Ozolai, – **"Par atbildību un valsts budžeta izdevumiem dalībai *Expo* izstādē"**.

Lūdzu, ministres kundze! Ir saņemta rakstiska atbildes, bet jautājuma uzdevēji nav apmierināti ar šo atbildi, vēlas saņemt arī papildu paskaidrojumus, uzdot papildu jautājumus. Bet vai jūs vēlaties sniegt arī vēl kādu papildu informāciju pirms jautājumu uzdošanas?

D.Reizniece-Ozola (ekonomikas ministre).

Komentāru.

Sēdes vadītāja. Jā, lūdzu!

D.Reizniece-Ozola. Labvakar, Lībiņas-Egneres kundze, un labvakar, deputāti! Man vispirms jāatvainojas par to, ka atbilde jums rakstiski ir iesniegta vēlāk, nekā bija lūgts no jūsu puses.

Vairāki apsvērumi.

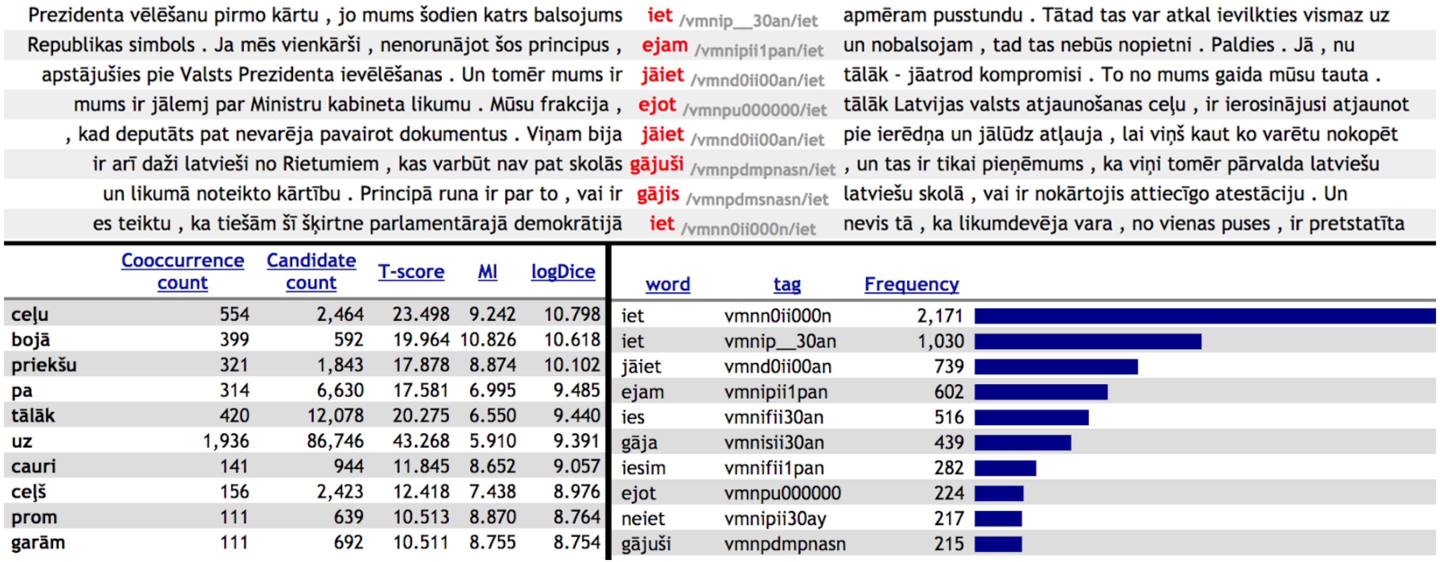
Viens. Mēs saņēmām ļoti vēlu to no premjerministres puses.

Un otrs. Šobrīd, kā jūs zināt, Ministru kabinets ir pieņēmis lēmumu pārtraukt Latvijas dalību *Expo* izstādē, un ir dots uzdevums Ekonomikas ministrijai gan lauzt līgumu, gan veikt iekšējas dienesta pārbaudes. Līdz ar to tās atbildes, kas tiek sniegtas, ir, nu, tādas diezgan piesardzīgas un varbūt nav pārāk izvērstas, lai pēcāk mums neradītu kādas saistošas juridiskas sekas.

Bonito corpus browser (NoSketch engine)

Text corpus user interface provides a powerful query system.

Queries can include words, lemmas, morphological tags and meta data.

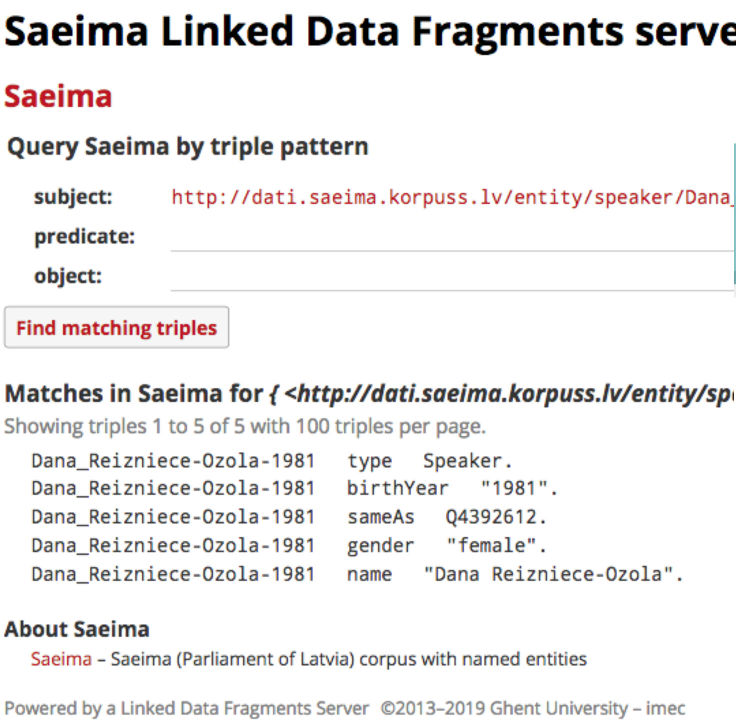


Prezidenta vēlēšanu pirmo kārtu , jo mums šodien katrs balsojums	iet /vmnip__30an/iet	apmēram pusstundu . Tātad tas var atkal ievilkties vismaz uz
Republikas simbols . Ja mēs vienkārši , nenorunājot šos principus ,	ejam /vmnipii1pan/iet	un nobalsojam , tad tas nebūs nopietni . Paldies . Jā , nu
apstājušies pie Valsts Prezidenta ievēlēšanas . Un tomēr mums ir	jāiet /vmnd0ii00an/iet	tālāk - jāatrod kompromisi . To no mums gaida mūsu tauta .
mums ir jālemj par Ministru kabineta likumu . Mūsu frakcija ,	ejot /vmnpu000000/iet	tālāk Latvijas valsts atjaunošanas ceļu , ir ierosinājusi atjaunot
, kad deputāts pat nevarēja pavairot dokumentus . Viņam bija	jāiet /vmnd0ii00an/iet	pie ierēdņa un jālūdz atļauja , lai viņš kaut ko varētu nokopēt
ir arī daži latvieši no Rietumiem , kas varbūt nav pat skolās	gājuši /vmnpdmpnasn/iet	, un tas ir tikai pieņēmums , ka viņi tomēr pārvalda latviešu
un likumā noteikto kārtību . Principā runa ir par to , vai ir	gājis /vmnpdmsnasn/iet	latviešu skolā , vai ir nokārtojis attiecīgo atestāciju . Un
es teiktu , ka tiešām šī šķirtne parlamentārajā demokrātijā	iet /vmnn0ii000n/iet	nevis tā , ka likumdevēja vara , no vienas puses , ir pretstatīta

	Cooccurrence count	Candidate count	T-score	MI	logDice
ceļu	554	2,464	23.498	9.242	10.798
bojā	399	592	19.964	10.826	10.618
priekšu	321	1,843	17.878	8.874	10.102
pa	314	6,630	17.581	6.995	9.485
tālāk	420	12,078	20.275	6.550	9.440
uz	1,936	86,746	43.268	5.910	9.391
cauri	141	944	11.845	8.652	9.057
ceļš	156	2,423	12.418	7.438	8.976
prom	111	639	10.513	8.870	8.764
garām	111	692	10.511	8.755	8.754

word	tag	Frequency
iet	vmnn0ii000n	2,171
iet	vmnip__30an	1,030
jāiet	vmnd0ii00an	739
ejam	vmnipii1pan	602
ies	vmnifii30an	516
gāja	vmnisii30an	439
iesim	vmnifii1pan	282
ejot	vmnpu000000	224
neiet	vmnipii30ay	217
gājuši	vmnpdmpnasn	215

http://dati.saeima.korpuss.lv/nosketch



Saeima Linked Data Fragments server

Saeima

Query Saeima by triple pattern

subject: http://dati.saeima.korpuss.lv/entity/speaker/Dana_

predicate:

object:

Find matching triples

Matches in Saeima for { <http://dati.saeima.korpuss.lv/entity/sp

Showing triples 1 to 5 of 5 with 100 triples per page.

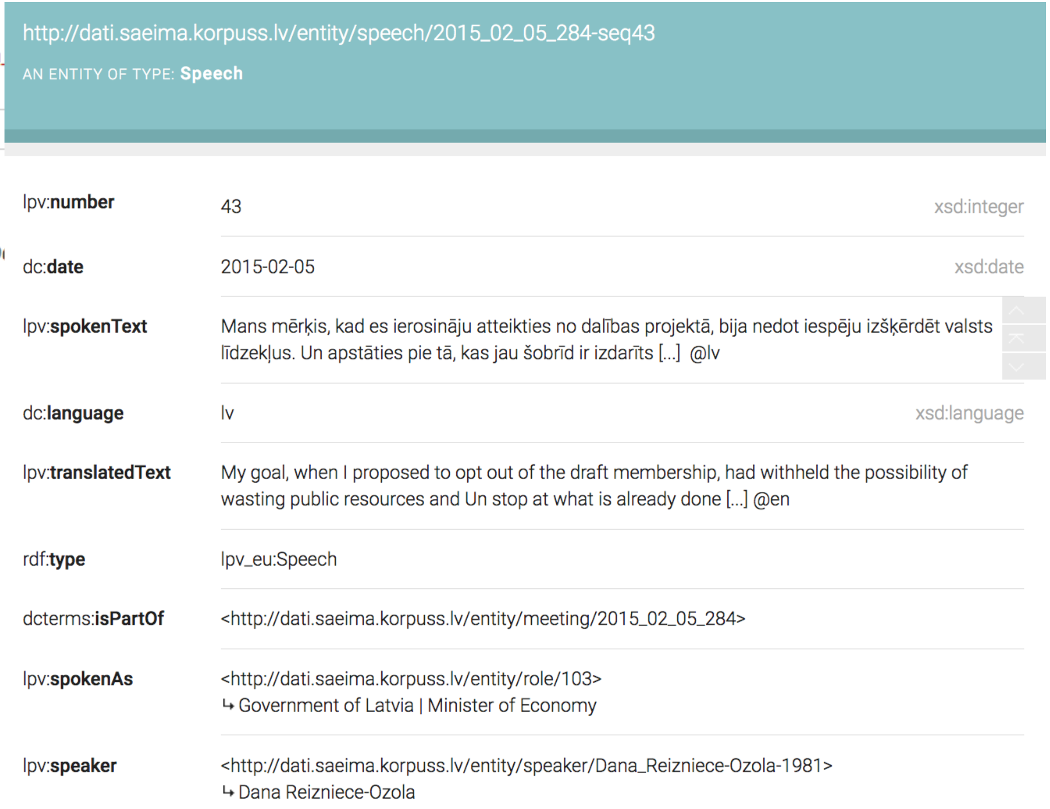
```
Dana_Reizniece-Ozola-1981   type   Speaker.
Dana_Reizniece-Ozola-1981   birthYear   "1981".
Dana_Reizniece-Ozola-1981   sameAs   Q4392612.
Dana_Reizniece-Ozola-1981   gender   "female".
Dana_Reizniece-Ozola-1981   name   "Dana Reizniece-Ozola".
```

About Saeima

Saeima – Saeima (Parliament of Latvia) corpus with named entities

Powered by a Linked Data Fragments Server ©2013–2019 Ghent University – imec

#LD
Linked Data Fragments



http://dati.saeima.korpuss.lv/entity/speech/2015_02_05_284-seq43

AN ENTITY OF TYPE: **Speech**

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	

- Triple Pattern Fragments
- Linked Data (LodView)
- RDF dump (Turtle RDF)

Named Entity Linking

Mentions of named entities (persons, organizations, locations) are linked to Wikidata knowledge base identifiers.

- Preprocessing: Wikidata entity aliases are extended with Latvian morphological inflections and automatically generated alternate labels (for person and organization names)
- Linking is performed by matching speech text fragments with entity labels (choosing the most specific match).

The Corpus of the Saeima contains 393 thousand mentions of approx. 3 thousand unique entities. 33% of speeches contain entity mentions.

Note: we also considered DBpedia but it had ~3 times less coverage than Wikidata.

```
type,source,session_name,session_type,
date,id,speaker,category,subcategory,
role,sequence,text,entities

[...]

speech,2015_02_05_284.txt,kārtējā sēde,kārtējā,
2015/02/05,2015_02_05_284.txt_seq33,Inese_Lībiņa-Egnere,
Saeimas amatpersonas,Priekšsēdētāja biedri,pb,33,
"Paldies.

Pirmo jautājumu vēlas uzdot deputāte Inguna Sudraba. Lūdzu,
ieslēdziet mikrofonu!","[[""Inguna Sudraba"", ""Inguna
Sudraba"", 46, ""http://www.wikidata.org/entity/Q4445523"",
""person""]]"
```

- Speech metadata (date, ID, ...)
- Information about the speaker
- Text (original)
- List of named entities (JSON)

Machine Translation

A subset of speeches (starting from 2015) was translated from Latvian to English using a neural machine translation system

- in RDF: lpv:translatedText property
- translation component from the SUMMA project - http://summa-project.eu/

Unreviewed machine-generated translation is provided for quantitative analysis and to aid searchability and understanding by international researchers

Text quality of automated translation is not sufficient for qualitative analysis

LinkedSaeima – Dataset

The current version of the dataset, published in May 2019, consists of approx. 4.9 million RDF triples. It is published according to Linked Data principles.

The dataset contains 497221 speeches (utterances)
from 1293 parliament meetings.

These speeches were given by 690 speakers in 162 different speaker roles and contain 392530 mentions of 2998 unique Wikidata entities.

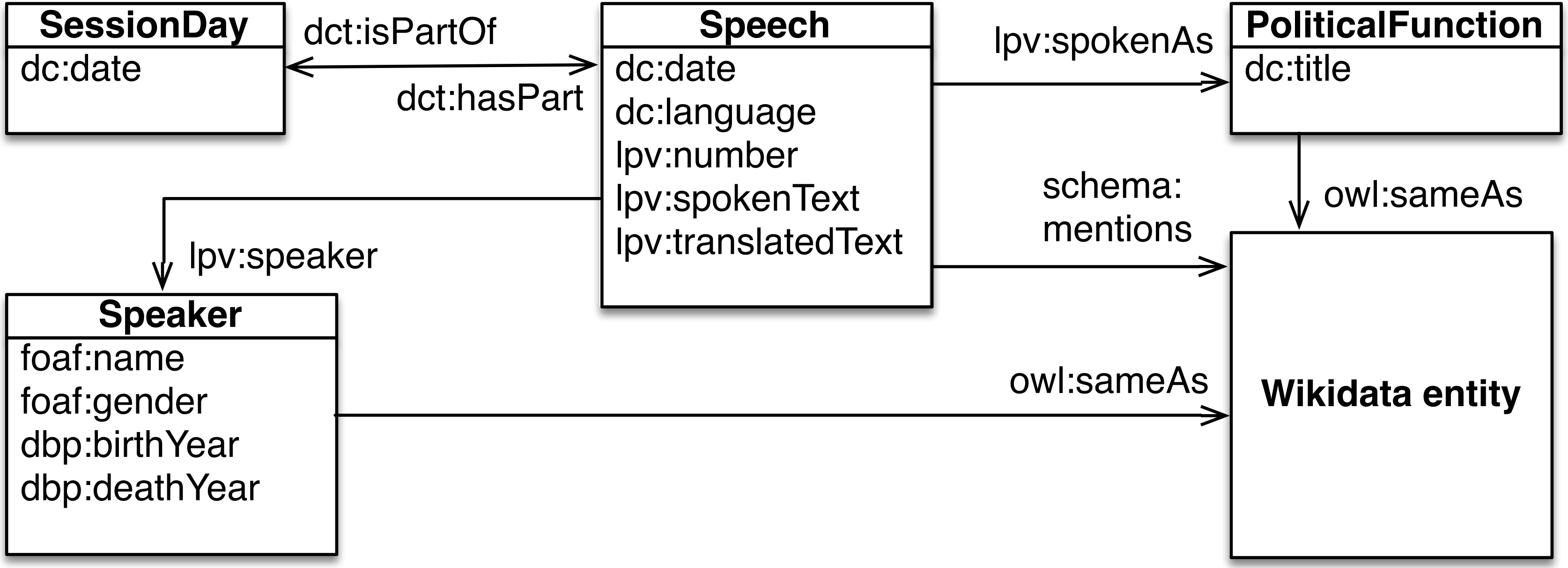
LinkedSaeima – Data Model

Main types of objects in the dataset:

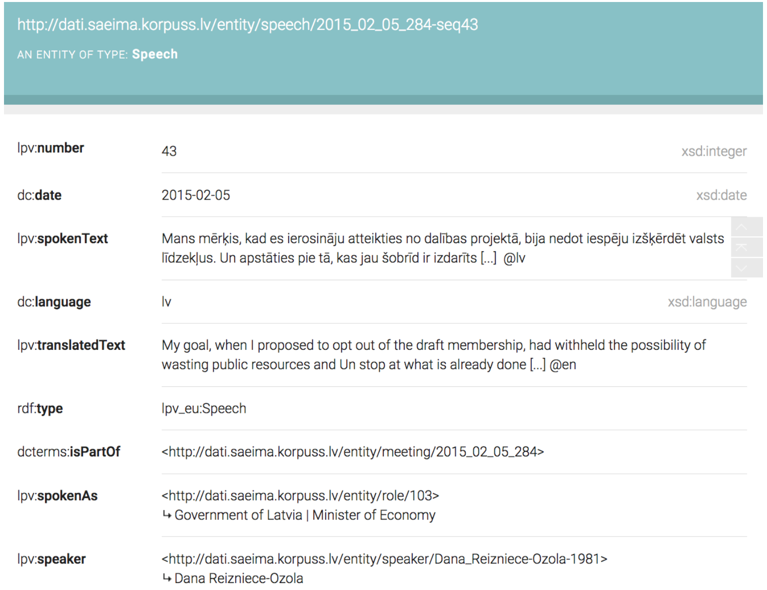
- Meeting – a top-level concept, represents one parliament meeting (a plenary) usually consisting of multiple Speeches
- Speech – an individual speech given at a Meeting by a particular Speaker acting in some Role (MP, minister, ...)
- Speaker – a person giving the speech
- Role – the role (e.g. Prime Minister) which the person represented
  when giving the Speech

[redacted]

Data Model (based on LinkedEP)



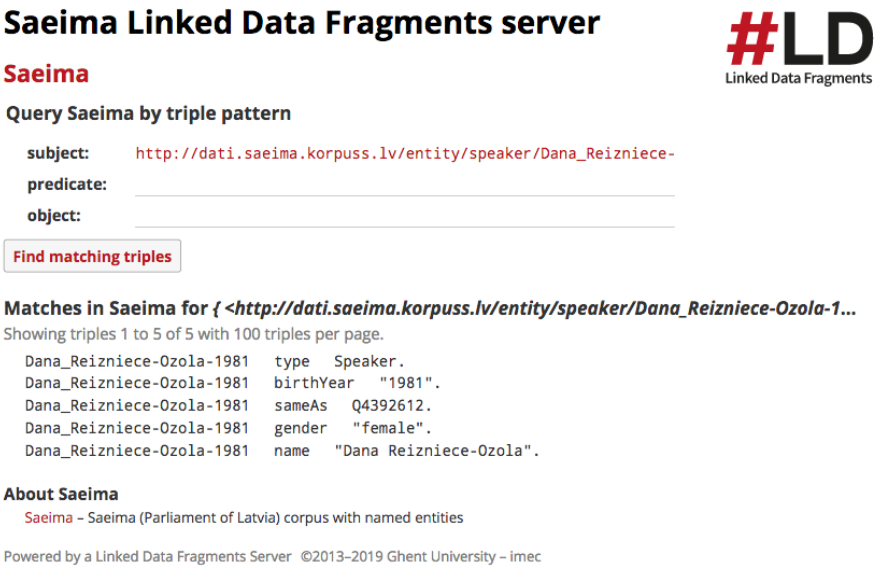
LinkedSaeima – Accessing the data



http://dati.saeima.korpuss.lv/entity/speech/2015_02_05_284-seq43

AN ENTITY OF TYPE: **Speech**

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	



Saeima Linked Data Fragments server

Saeima

Query Saeima by triple pattern

subject: http://dati.saeima.korpuss.lv/entity/speaker/Dana_Reizniece-

predicate:

object:

Matches in Saeima for *{ <http://dati.saeima.korpuss.lv/entity/speaker/Dana_Reizniece-Ozola-1...*

Showing triples 1 to 5 of 5 with 100 triples per page.

```
Dana_Reizniece-Ozola-1981  type   Speaker.
Dana_Reizniece-Ozola-1981  birthYear   "1981".
Dana_Reizniece-Ozola-1981  sameAs   Q4392612.
Dana_Reizniece-Ozola-1981  gender   "female".
Dana_Reizniece-Ozola-1981  name   "Dana Reizniece-Ozola".
```

About Saeima

Saeima – Saeima (Parliament of Latvia) corpus with named entities

Powered by a Linked Data Fragments Server ©2013–2019 Ghent University – imec

LinkedSaeima entity information
(LodView)

LinkedSaeima
Triple Pattern Fragments server

[redacted]

LinkedSaeima – URI patterns

Type	**URI pattern**
Speech	`/entity/speech/2015_02_05_284-seq43`
Speaker	`/entity/speaker/Dana_Reizniece-Ozola-1981`
Role	`/entity/role/103`
SessionDay	`/entity/meeting/2015_02_05_284`

```
@base          <http://dati.saeima.korpuss.lv/entity/> .

<speech/2015_02_05_284-seq41>
    a                    lpv_eu:Speech ;
    dc:date              "2015-02-05"^^xsd:date ;
    dc:language          "lv"^^xsd:language ;
    dcterms:isPartOf     <meeting/2015_02_05_284> ;
    lpv:number           41 ;
    lpv:speaker          <speaker/Inese_Libina-Egnere-1977> ;
    lpv:spokenAs         <role/122> ;
    lpv:spokenText       "Tad pāriesim pie papildu jautājumu uzdošanas.\nTā
kā jautājuma iesniedzēju šeit nav, uzreiz pārējie deputāti, kas ir klāt, var
izmantot iespēju pavisam kopā uzdot trīs papildu jautājumus.\nLūdzu, sāksim
ar deputāti Ingunu Sudrabu.\nIeslēdziet, lūdzu, mikrofonu deputātei
Sudrabai!"@lv ;
    lpv:translatedText   "Let us then move on to the supplementary questions,
as the questioner is not here, the other straight Members who are present may
take the opportunity to put three additional questions together at Please,
please click on the MEP's Ingu silver printable."@en ;
    schema:mentions      <http://www.wikidata.org/entity/Q4445523> .
```

LinkedSaeima – Innovation

Adding named entity information pointing to Wikidata entity URIs.

- schema:mentions property

"Materialization" of speaker Roles by giving them URI identifiers and linking to Wikidata entity URIs.

- owl:sameAs property

Linking speaker entities (MPs, ...) to their Wikidata URIs,
to make it easier to link datasets to one another.

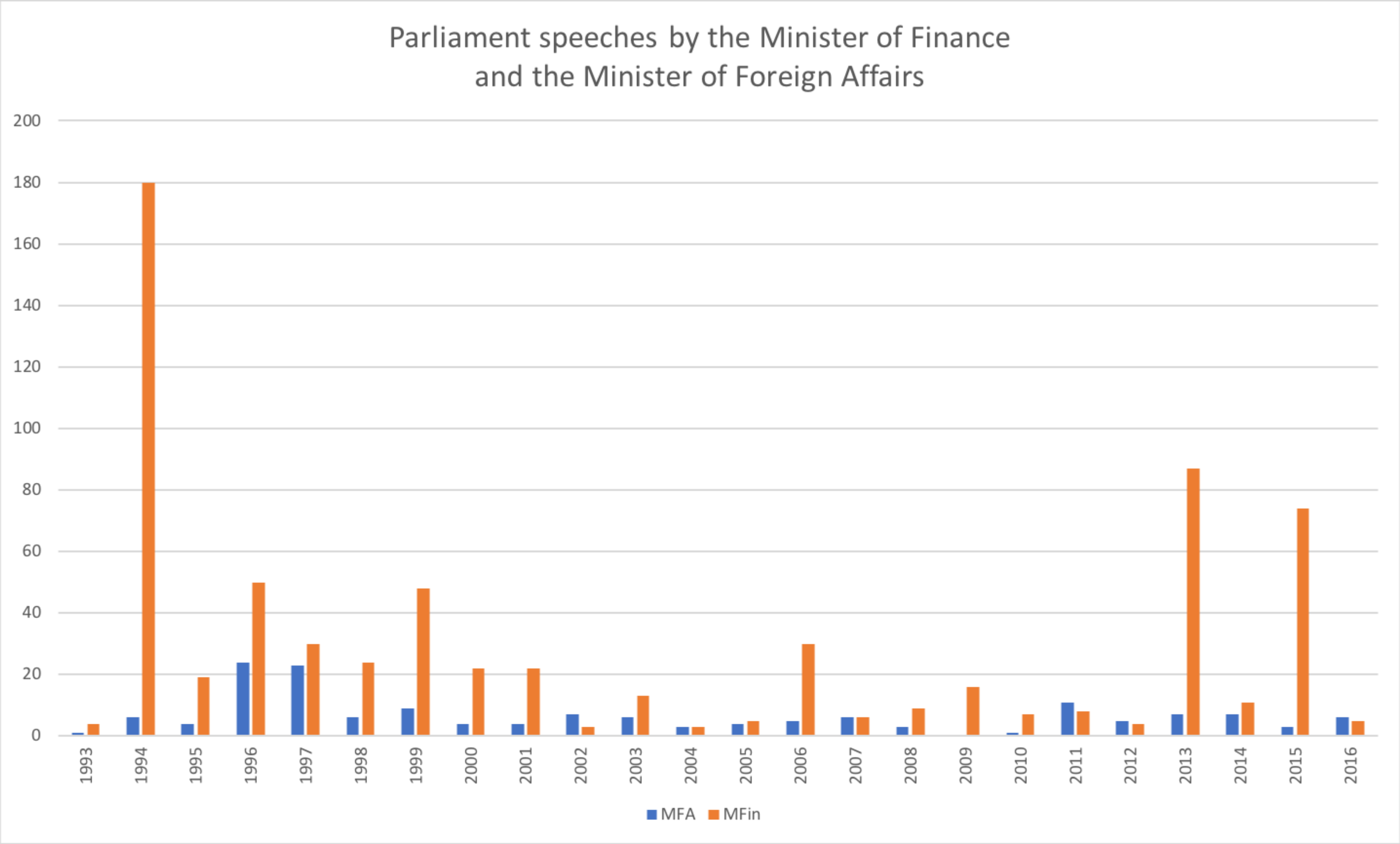
- owl:sameAs property

Using the Data

```
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
```

Yearly statistics of speeches
by the Minister of Foreign Affairs



Parliament speeches by the Minister of Finance and the Minister of Foreign Affairs

Conclusions

- Latvia's parliamentary debate corpus, enriched with additional information (named entities, automatic translation, ...).
- LinkedSaeima – a Linked Data representation of the Saeima speech corpus, with links to Wikidata entities. Data model based on the LinkedEP project.
- A rich source of information about Latvia's parliamentary debates, offering new ways how to explore this information.
- Future work:
  - Standardization of parliamentary corpora and datasets.
  - Adding voting data, improving linking to other datasets.

http://dati.saeima.korpuss.lv/

uldis.bojars@gmail.com