ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora

Miyazaki, 7 May 2018

Annotation of the Corpus of the Saeima with Multilingual Standards

Roberts Darģis, Ilze Auziņa, Uldis Bojārs,
Pēteris Paikens, Artūrs Znotiņš

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

Mākslīgā intelekta laboratorija
LU MII

LATVIJAS UNIVERSITĀTE
ANNO 1919
UNIVERSITY OF LATVIA

Motivation

- Current comparative research of parliamentary debate is not sufficiently facilitated
- Augmenting such corpora with extra layers according to multilingual standards would help

- Annotation Layers
  - English translation
  - Named entities
  - Morphological and syntactical annotations
- Available datasets
  - Linked Data
  - Universal Dependencies
  - Bonito corpus browser

The Corpus of the Saeima

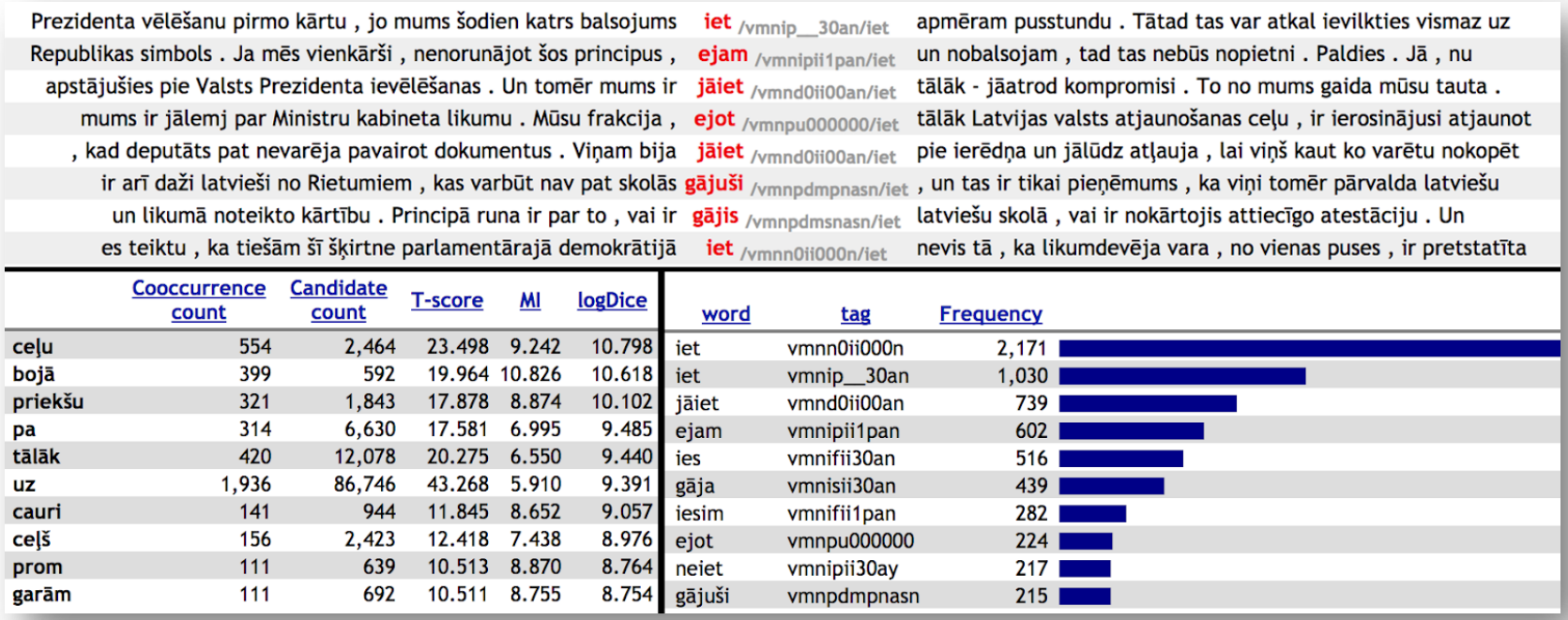
- Saeima has been the name of the parliament of the Republic of Latvia since 1922
- The Saeima is composed of 100 members of parliament and it is elected for a term of four years
- The Corpus of the Saeima includes transcriptions of parliamentary debate from 7 parliamentary terms (5th–12th), covering years 1993–2017
- The transcriptions of the Corpus of the Saeima contain 38 million tokens, 497 thousand utterances and 468 speakers

Morphological and Syntactical Annotations

- The annotations contain lemma, part of speech, morphological features and syntactic dependencies according to the Universal Dependencies standard format
- Texts are automatically tokenized, lemmatized and morphologically analyzed and tagged using CMM based tagger
- Syntactic dependencies are inferred by neural transition-based dependency parser trained on Latvian Universal Dependencies corpus version 2.1

Bonito corpus browser (NoSketch engine)

- The interface provides powerful corpus query system. Query 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

Universal Dependencies (CoNLL-U)

- Automatic tokenization, morphological and syntactic annotations are published in CoNLL-U data format

```
# newdoc id = 2016_03_31_355.txt_seq17
# newpar id = 2016_03_31_355.txt_seq17-p1
# sent_id = 2016_03_31_355.txt_seq17-p1s1
# text = Turpinām ar iesniegtajām izmaiņām Saeimas Prezidija apstiprinātajā
sēdes darba kārtībā.
 1  Turpinām      turpināt     _  vmnipt31pan      _  0  root   _  _
 2  ar            ar           _  sppd             _  4  case   _  _
 3  iesniegtajām  iesniegt     _  vmnpdfpdpsyp     _  4  amod   _  _
 4  izmaiņām      izmaiņa      _  ncfpd4           _  1  iobj   _  _
 5  Saeimas       saeima       _  ncfsg4           _  6  nmod   _  _
 6  Prezidija     prezidijs    _  ncmsg1           _  8  nmod   _  _
 7  apstiprinātajā apstiprināt _  vmnpdfslpsyp    _  8  amod   _  _
 8  sēdes         sēde         _  ncfsg5           _  10 nmod   _  _
 9  darba         darbs        _  ncmsg1           _  10 nmod   _  _
10  kārtībā       kārtība      _  ncfsl4           _  1  obl    _  SpaceAfter=No
11  .             .            _  zs               _  1  punct  _  _
```

Machine Translation to English

- The speeches from Latvian are translated to English using a neural machine translation system
- The unreviewed machine-generated translation is provided for quantitative analysis and to aid searchability and understanding for international researchers
- The text quality of automated translation is lacking, so for qualitative analysis a professional translator should be used

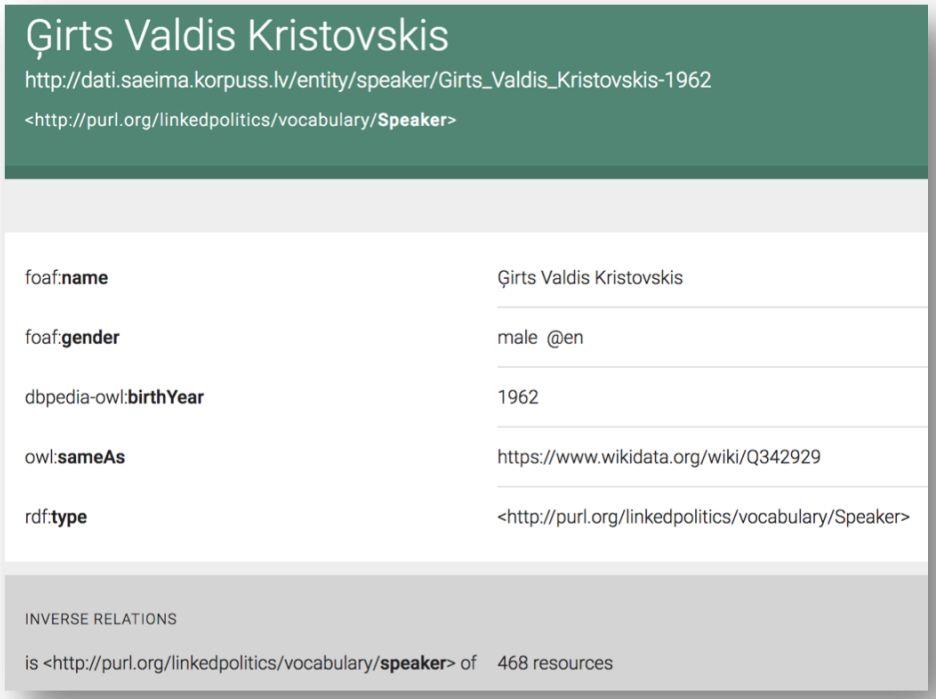
Named Entities

- We used the structured Wikidata information extracts as the entity knowledge base. The Wikidata entity alias information is extended with Latvian morphological inflections and automatically generated variants for people and organization names
- In the Corpus of the Saeima we identified 393 thousand mentions of 3 thousand unique entities. 165 thousand out of 497 thousand utterances contained entity mentions

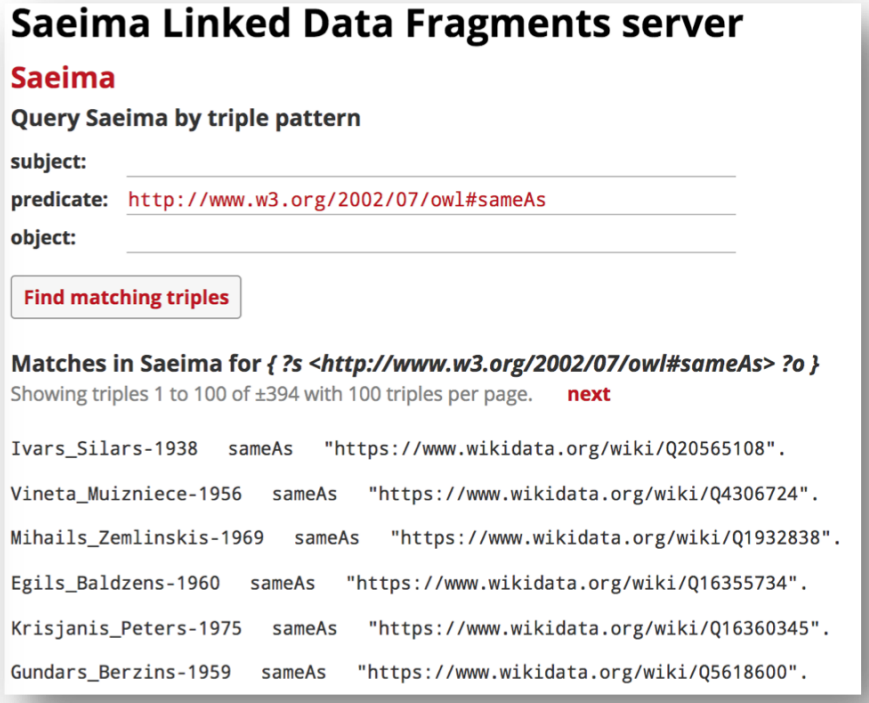
LinkedSaeima I – structure

- Structured information about parliamentary debates is represented using Resource Description Framework (RDF), according to the Linked Data principles
- The types of objects in the LinkedSaeima dataset are:
  - Meeting – a top-level concept representing one parliament meeting (a plenary) usually consisting of multiple Speeches
  - Speech – an individual speech given at a Meeting by a particular Speaker in some Role
  - Speaker – a person giving a speech
  - Role – a role (e.g. Prime Minister) which the person represented when giving a Speech

LinkedSaeima II – interfaces



Screenshot of a
LinkedSaeima entity in LodView



Screenshot of the LinkedSaeima
triple pattern fragments server

LinkedSaeima index page - *http://dati.saeima.korpuss.lv*

LinkedSaeima III – innovation

- Main innovation of this dataset, relative to the LinkedEP project:
  - Addition of named entity information, pointing to corresponding Wikidata URI indentifiers
  - “Materialization” of speaker Roles, by giving them URI identifiers and linking them to Wikidata URI identifiers
  - Manually linking speakers to Wikidata URI, to make it easier to conduct a inter-corpora research

Conclusions

- The new annotation levels and its Linked Data representation will widen the applications for *The Corpus of the Saeima*
- Future work includes improvements to annotation tools, and extending the LinkedSaeima dataset
- We'd like to call upon this research community to pursue open, common NLP data standards to enable multilingual comparative research

Thank you!

Resource and description: http://dati.saeima.korpuss.lv

NATIONAL DEVELOPMENT PLAN 2020
EUROPEAN UNION
European Regional Development Fund
INVESTING IN YOUR FUTURE
Mākslīgā intelekta laboratorija
LU MII
LATVIJAS UNIVERSITĀTE
ANNO 1919
UNIVERSITY OF LATVIA