# **CLARIN Café**

Join Our Parliamentary-flavoured Coffee: ParlaMint

## The agenda

14:10–14:30

13:00–13:10 ParlaMint in a Nutshell: New Comparable Parliamentary Corpora in the Pandemic Fra Maciej Ogrodniczuk, Petya Osenova 13:10–13:25 How to standardize the Parliamentary Data and Metadata in a Multilingual Context Tomaž Erjavec, Andrej Pančur 13:25–13:40 How to make the Parliamentary Data even better discoverable: ahead with the Linguistic Processing Nikola Ljubešić 13:40–13:55 How to extract useful parliamentary knowledge: Concordancing the Parliament Tomaž Erjavec 13:55–14:10 Join Us as a Contributor and/or as a User: Possibilities, Potential and Perspectives Maciej Ogrodniczuk, Petya Osenova

QA and Discussion

# ParlaMint in a Nutshell:

# New Comparable Parliamentary Corpora in the Pandemic Era

Maciej Ogrodniczuk and Petya Osenova

### What is ParlaMint?



A mini-project supported by CLARIN-ERIC

**Budget:** 98,000 €

**Duration:** Jul 1 2020 – May 30 2021

Motivation: Parliamentary data directly corresponds to the most recent events with global impact on *human health, social life* and *economics* such as the current COVID-19 pandemic.

Goal: Provide resources and tools for focused observations on trends, opinions, decisions on *lockdowns* and *restrictive measures* as well as on *the consequences* with respect to health, medical care systems, employment, etc. during pandemic times.

## **How is ParlaMint implemented?**



Phase 1 (July 2020 – Sep 2020):

The corpus of *Bulgarian, Croatian, Polish* and *Slovene* parliamentary speeches was created with two parts marked:

- COVID-19 subcorpus (Nov 2019 2020)
- reference subcorpus (2015 Nov 2019).

Phase 2 (Dec 2020 - May 2020):

Parliaments for 6 more countries will be created according to the methodology established in Phase 1.

(Stay for the last part for more information how to participate!)

### **ParlaMint workflow**

- 1. Getting the parliamentary data and metadata
- 2. Converting them into the Parla-CLARIN schema
- 3. Validation (formal and qualitative)
- Linguistic annotation with syntactic structures of Universal Dependencies and Named Entities.
- 5. Making the corpora available through concordancers and Parlameter.
- 6. Next: Adding more parliaments
- Next: Building use cases in Political Sciences and Digital Humanities based on the corpus data.

### **Meet the participants:)**

- Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences (also: Michał Rudolf)
- Petya Osenova, Institute for Information and Communication Technologies, Bulgarian Academy of Sciences (also: Kiril Simov)
- Tomaž Erjavec, Jožef Stefan Institute, Ljubljana (also: Darja Fišer and Kristina Pahor de Maiti)
- Andrej Pančur, Institute of Contemporary History, Ljubljana
- Nikola Ljubešić, Jožef Stefan Institute, Ljubljana
- Filip Muki Dobranić, Parlameter.org (also: Tomaž Kunst)
- Ruben Ros, Luxembourg University

### What is available: First phase results

- Parla-CLARIN schema (https://github.com/clarin-eric/parla-clarin)
- Corpora to download from CLARIN.SI repository (http://hdl.handle.net/11356/1345)
- Corpora in noSketch Engine (https://www.clarin.si/noske/parlamint.cgi)
- Corpora in KonText (https://www.clarin.si/kontext/)
- Ongoing Now: Corpora in Parlameter (https://parlameter.org/)
- Know-how on dealing with parliamentary data!

# How to standardize

# the Parliamentary Data and Metadata in a Multilingual Context

Tomaž Erjavec and Andrej Pančur

### **Background**

- CLARIN organised the <u>ParlaCLARIN workshop</u> at LREC 2018 and made parliamentary corpora as one of its <u>Key resource families</u>
  - many corpora exist, but are encoded in many different ways, limiting interchange

- CLARIN Type B workshop: <u>ParlaFormat</u>, May 23-24, 2019: Propose a recommendation for encoding parliamentary corpora and get community feedback
- Parla-CLARIN: a TEI based encoding for parliamentary corpora
- Maintained on GitHub: <a href="https://github.com/clarin-eric/parla-clarin">https://github.com/clarin-eric/parla-clarin</a>
- So far a simple customisation of TEI, using the base module for spoken language corpora
- The most important part is the documentation: <a href="https://clarin-eric.github.io/parla-clarin/">https://clarin-eric.github.io/parla-clarin/</a>

### Parla-CLARIN

### The recommendations attempt to take into account:

- Structure: legislative periods, sessions, topics, speeches, transcription variants
- Metadata: mandates, titles, parliamentary bodies, locations, dates and times
- Speakers: date of birth, sex, education, party membership, links to external resources
- Political parties: name(s), history, relations

### Significant work remains to be done:

- Stricter schema
- Better examples
- Scripts for down-conversion
- Community development



### ParlaMint schema

Bottom-up procedure: started with Parla-CLARIN and then unified the encoding

- The result is the 4-language ParlaMint 1.0 corpus, with a quite uniform encoding: <a href="http://hdl.handle.net/11356/1345">http://hdl.handle.net/11356/1345</a>
- We also made a bespoke XML schema, which significantly constrains the Parla-CLARIN encoding, and is available at the repository entry
- The easiest way to encode new ParlaMint corpora:
  - Take one of the existing ParlaMint corpora as an example
  - Validate with the ParlaMint schema
  - Get in touch with Tomaž Erjavec in case things are unclear

### **Further work**

There are still things that could be done better in the ParlaMint encoding:

- Better encoding of transcriber comments
- Further harmonisation of the legislation and other taxonomies
- Integration with Parla-CLARIN

Hopefully, we will do this in the second phase of the project.



# How to make the Parliamentary Data even better discoverable:

ahead with the Linguistic Processing

## **Automatic linguistic processing**

Unified! linguistic processing performed in an automated manner with tools based on machine learning

- Levels of linguistic processing:
  - Tokenization and sentence splitting
  - Lemmatization
  - Morphosyntactic annotation
  - Dependency parsing
  - Named entity recognition

## **Example of linguistic processing result**

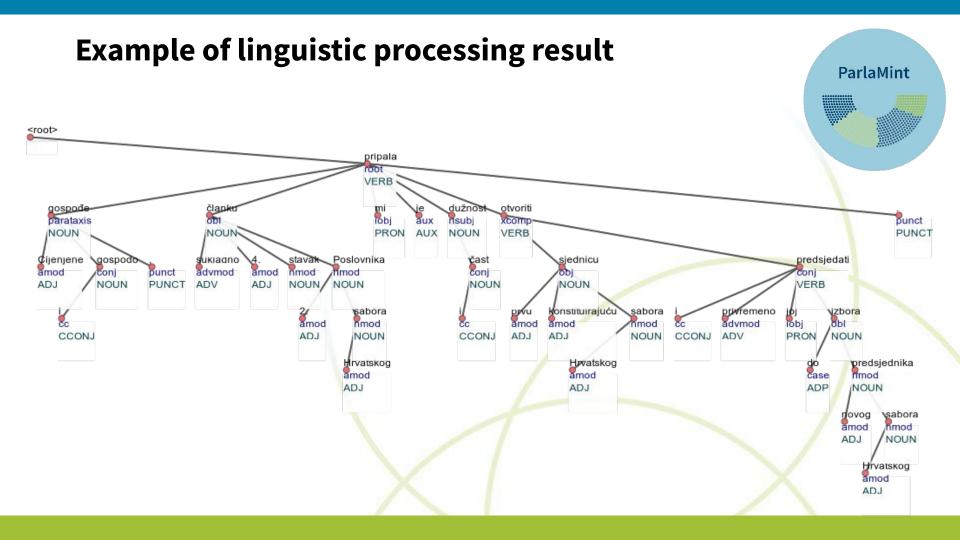


### First sentence of the Croatian corpus

| 1  | Cijenjene  | cijeniti  | ADJ   | 2  | amod   | NER=O     |
|----|------------|-----------|-------|----|--------|-----------|
| 2  | gospođe    | gospođa   | NOUN  | 14 | nsubj  | NER=O     |
| 3  | i          | i         | CCONJ | 4  | СС     | NER=O     |
| 4  | gospodo    | gospod    | NOUN  | 2  | conj   | NER=O     |
| 5  | ,          | ,         | PUNCT | 7  | punct  | NER=O     |
| 6  | sukladno   | sukladno  | ADV   | 7  | advmod | NER=O     |
| 7  | članku     | članak    | NOUN  | 2  | appos  | NER=O     |
| 8  | 4.         | 4.        | ADJ   | 9  | amod   | NER=O     |
| 9  | stavak     | stavak    | NOUN  | 7  | appos  | NER=O     |
| 10 | 2.         | 2.        | ADJ   | 11 | amod   | NER=O     |
| 11 | Poslovnika | poslovnik | NOUN  | 9  | nmod   | NER=O     |
| 12 | Hrvatskog  | hrvatski  | ADJ   | 13 | amod   | NER=B-ORG |
| 13 | sabora     | sabor     | NOUN  | 11 | nmod   | NER=I-ORG |

Cijenjene gospođe i gospodo, sukladno članku 4. stavak 2. Poslovnika Hrvatskog sabora...

Distinguished ladies and gentlemen, in accordance with article 4, paragraph 2 of the Rules of Procedure of the Croatian Parliament...



## For the technically savvy

- Bulgarian, Croatian, Slovenian were processed with the CLASSLA pipeline (<a href="https://pypi.org/project/classla/">https://pypi.org/project/classla/</a>), spin-off of Stanford Stanza with larger training datasets (beyond treebanks), inflectional lexicons, added NER
  - state-of-the-art for all three languages, morphosyntax ~97(/100), lemmatization ~99,
     dependency syntax ~93, named entities ~98, (Bi)LSTM technology

**ParlaMint** 

Polish was processed with inflectional analyser <u>Morfeusz</u>, morphosyntactic tagger <u>Concraft</u>, name entities detector <u>Liner</u> and dependency parser <u>COMBO</u>, similar numbers as for the tools above

# Limitations of automatic linguistic processing!



Automatic linguistic processing is as good as the data the tools were trained on

All of the presented annotation layers have mistakes. From the previous slides...

```
gospodo gospoda NOUN
```

```
sukladno [članku 4.] [stavak 2.] [Poslovnika [Hrvatskog sabora]]
```

# **Possible ways forward**

- Extending training data with domain-specific data
  - "Cijenjene gospođe i gospodo"
  - "sukladno članku 4. stavak 2. Poslovnika Hrvatskog sabora"
- Named entity recognition extended to named entity linking ("Hrvatskog sabora" being labeled with the ID of the "Croatian parliament" entity)
- Coreference resolution ("them" referring to the opposition)
- Sentiment / emotion analysis
- Extending with audio and video recordings (vocal features, gestures)

# How to extract useful parliamentary knowledge:

**Concordancing the Parliament** 

### **Concordancers**

- Corpus analysis tools, where the basic functionality is concordances, i.e. showing words in context
- Best known commercial web-based concordancer is Sketch Engine
- Lexical computing also makes available an open source version (with somewhat reduced but still powerful functionality) noSketch Engine:
  - manatee back-end
  - bonito front-end (now no longer maintained)
  - new front-end: Crystal
- The Czech National Corpus group produced KonText:
  - uses manatee as the back-end but different front-end
  - <u>LINDAT/CLARIAH-CZ</u> made a fork of KonText, which supports AAI
- CLARIN.SI has an installation of noSketch Engine / Bonito and CLARIAH-CZ KonText
  - https://www.clarin.si/info/concordances/
  - both offer the same ~50 corpora



## noSketch Engine / KonText Functionality

- Support for large corpora annotated with structural (e.g. speech, sentence, name) and positional (lemma, PoS, morphology, syntactic head) information
- Powerful corpus query language CQL:
  - logical expressions for tokens, e.g. [lemma="love" & pos="NOUN"]
  - token sequences, e.g. [lemma="love"][word="to"]?[]+[lemma="work"]
  - regular expressions, e.g. "l[aeiou]v.+"
  - structural constraints, e.g. ".+ets" within <name>
- Display of
  - concordances
  - frequency lists
  - keyword lists (noSketch Engine only)
  - collocations
- Saving and searching subcorpora, personalising the display
- Personal space (KonText only)
- REST interface

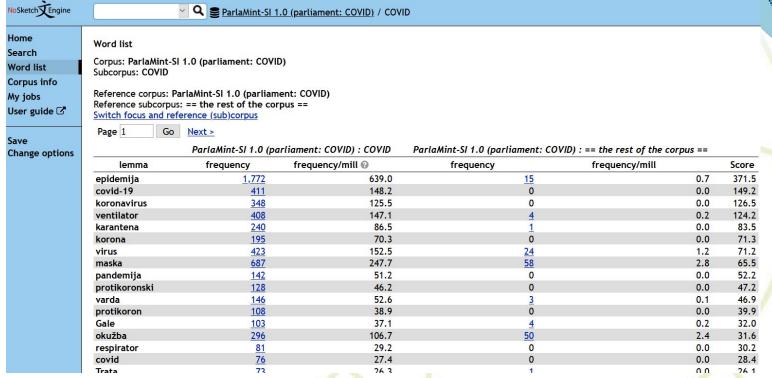
## **Analysing ParlaMint**

- ParlaMint 1.0 converted from TEI to vertical format for concordancers
  - Flattening the structure and renaming of attributes
  - Vertical files and conversion XSLT included in corpus distribution
- Publicly available, also through the dedicated entry point <a href="https://www.clarin.si/noske/parlamint.cgi/">https://www.clarin.si/noske/parlamint.cgi/</a>



# **COVID Period Key Lemmas in ParlaMint-SI 1.0**





## Male vs. Female Key Lemmas



| ParlaMint-SI 1.0 ( | parliament: COVID | ): Male |
|--------------------|-------------------|---------|
|--------------------|-------------------|---------|

|                        | The state of the s |                  |  |  |
|------------------------|--|------------------|--|--|
| lemma                  | frequency  | frequency/mill @ |  |  |
| lev <mark>i</mark> čar | 364  | 21.3             |  |  |
| navzoč                 | 4,326  | 252.9            |  |  |
| penez                  | 167  | 9.8              |  |  |
| esklopa                | 212  | 12.4             |  |  |
| Iran                   | 260  | 15.2             |  |  |
| gorica                 | <u>168</u>   | 9.8              |  |  |
| počenjati              | 182  | 10.6             |  |  |
| islam                  | <u>180</u>   | 10.5             |  |  |
| levičarski             | <u>163</u>   | 9.5              |  |  |
| Palestina              | <u>111</u>   | 6.5              |  |  |
|                        |  |                  |  |  |

ParlaMint-SI 1.0 (parliament: COVID) : Female

|              | ,                          |      |  |  |  |
|--------------|----------------------------|------|--|--|--|
| lemma        | frequency frequency/mill ② |      |  |  |  |
| Medical      | 107                        | 16.7 |  |  |  |
| KB           | <u>56</u>                  | 8.7  |  |  |  |
| pacientov    | 210                        | 32.7 |  |  |  |
| enaročanje   | <u>75</u>                  | 11.7 |  |  |  |
| ennapotnica  | <u>42</u>                  | 6.5  |  |  |  |
| žilen        | <u>417</u>                 | 64.9 |  |  |  |
| psihiatričen | <u>69</u>                  | 10.7 |  |  |  |
| roditeljski  | <u>80</u>                  | 12.5 |  |  |  |
| kardiolog    | <u>80</u>                  | 12.5 |  |  |  |
| zaznamba     | <u>100</u>                 | 15.6 |  |  |  |
| rejništvo    | 206                        | 32.1 |  |  |  |
| opornica     | <u>422</u>                 | 65.7 |  |  |  |
| posvojitev   | 328                        | 51.1 |  |  |  |
|              |                            |      |  |  |  |

# Join Us as a Contributor and/or as a User:

Possibilities, Potential and Perspectives

Maciej Ogrodniczuk and Petya Osenova

# If you want to join us as a contributor

 We invite proposals to add parliamentary corporal for additional countries to the ParlaMint collection

Call schedule:

**26 Oct 2020:** Call issued

16 Nov 2020: Submission deadline

20 Nov 2020: Results announced

**1 Dec 2020:** Projects start

30 Mar 2021: Delivery of the corpora



# If you want to join us as a contributor

### We expect:

- Obtaining data for the COVID-19 and reference parts of the corpus
- Conversion of the data into the ParlaMint format
- Linguistic processing of the corpus with Universal Dependencies, preferably including a suitable NER module
- Producing documentation on the provided corpus

### We provide:

- Dedicated guidelines on how to prepare the data
- The already existing corpora as models and the ParlaMint Schema:
- Upload to the concordancers
- Upload to Parlameter



### **More details**

Funding: 5,000 Euro per project

Duration: Dec 1, 2020 – Mar 31, 2021 (4 months)

### Assessment criteria:

- Qualifications of the team involved
- Status of the available and/or accessible parliamentary corpora
- Diversity of parliaments and languages
- Potential with respect to the project goals



## If you want to join as a User

### Three showcases are envisaged:

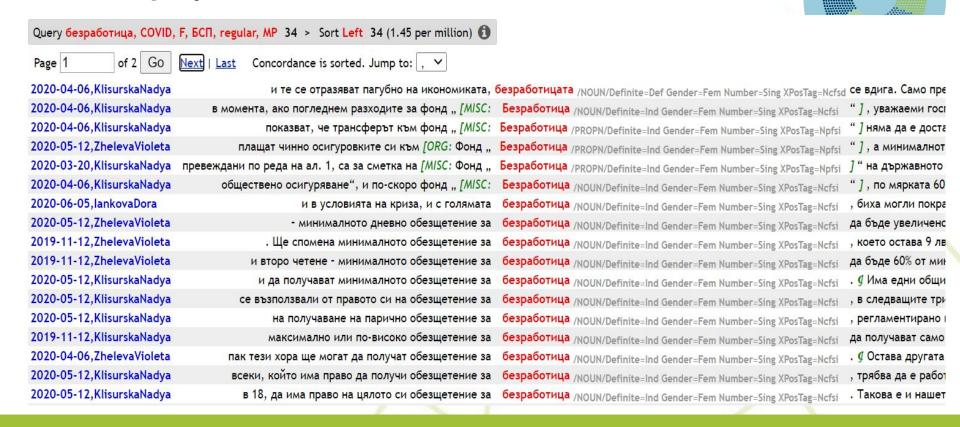
- Parlameter showcase: graphical exploration of the corpus suited for investigations by political scientists, journalists and active citizens
- Extending the CLARIN tutorial "Voices of the Parliament":
   how corpora can be used to investigate language use and communication
   practices in a specialised socio-cultural context of political discourse.
- 3. A DH-related showcase prepared by our digital historian.

But you do not have to wait for all this to come.

Start exploring the data **NOW** through the concordancers!!! :)



# The utterances of women, MPs, from socialist opposition: unemployment





### Collocation candidates

| Vi.                      | Cooccurrence<br>count | Candidate<br>count | T-score | MI     | <u>logDice</u> |
|--------------------------|-----------------------|--------------------|---------|--------|----------------|
| <u>Р   N</u> обезщетение | 16                    | 2,416              | 3.999   | 12.155 | 7.741          |
| <u>Р   N минимален</u>   | 7                     | 4,381              | 2.643   | 10.104 | 5.699          |
| Р   N девет              | 4                     | 3,606              | 1.997   | 9.577  | 5.170          |
| <u>Р   N</u> оставам     | 3                     | 3,268              | 1.729   | 9.304  | 4.895          |
| Р   N дневен             | 3                     | 3,405              | 1.729   | 9.245  | 4.837          |
| <u>Р   М</u> фонд        | 6                     | 9,863              | 2.443   | 8.710  | 4.312          |
| P   N xopa               | 3                     | 25,862             | 1.710   | 6.320  | 1.924          |
| Р   N свой               | 4                     | 36,715             | 1.973   | 6.229  | 1.834          |

# The utterances of women, MPs, from ruling party: unemployment







#### Collocation candidates

|                            | Cooccurrence<br>count | Candidate<br>count | T-score | MI     | <u>logDice</u> |
|----------------------------|-----------------------|--------------------|---------|--------|----------------|
| <u>Р   N</u> смекчаване    | 4                     | 31                 | 1.999   | 16.397 | 10.955         |
| <u>Р   N</u> обезщетение   | 9                     | 2,416              | 2.998   | 11.283 | 6.910          |
| <u>Р   N</u> паричен       | 3                     | 1,020              | 1.731   | 10.942 | 6.541          |
| Р I N фонд                 | 13                    | 9,863              | 3.601   | 9.784  | 5.427          |
| Р   N извънреден           | 3                     | 2,752              | 1.729   | 9.510  | 5.140          |
| <u>Р   № обстоятелство</u> | 3                     | 3,007              | 1.729   | 9.382  | 5.014          |
| P   N puck                 | 4                     | 4,192              | 1.996   | 9.318  | 4.954          |
| <u>Р   N</u> намаляване    | 3                     | 3,286              | 1.729   | 9.254  | 4.887          |
| <u>Р   N</u> осигуряване   | 5                     | 7,899              | 2.230   | 8.726  | 4.368          |
| <u>Р   N</u> получавам     | 3                     | 5,666              | 1.727   | 8.468  | 4.107          |
| <u>Р   N</u> обществен     | 5                     | 14,155             | 2.226   | 7.884  | 3.529          |
| <u>Р   N</u> размер        | 3                     | 9,965              | 1.723   | 7.654  | 3.297          |
| <u>Р   № цел</u>           | 4                     | 14,509             | 1.989   | 7.527  | 3.171          |
| <u>Р   N</u> бюджет        | 3                     | 22,086             | 1.712   | 6.506  | 2.151          |

# **QA and Discussion**

# Thank you and see you at:

https://www.clarin.eu/content/parlamint