ParlaMint II: The show must go on

Maciej Ogrodniczuk, Petya Osenova, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Çagrı Çöltekin, Matyáš Kopp, Katja Meden

20.06.2022, ParlaCLARIN III, LREC 2022

Plan of the Talk

- Introduction
- Schema and Metadata Improvements
- Corpus Expansion
- Corpus Enrichment
- Engagement Activities
- Beyond ParlaMint II

ParlaMint: the first CLARIN flagship project

- Financially supported by CLARIN
- ParlaMint I (July 2020 – May 2021) – see (Erjavec et al., 2022)
  - created and made available corpora for 17 parliaments
  - started to use them in training and research
- ParlaMint II (December 2021 – May 2023)
  - upgrading the XML schema and validation
  - extending existing corpora and enhancing the corpora with additional metadata
  - adding corpora for new parliaments
  - improving the usability of the corpora

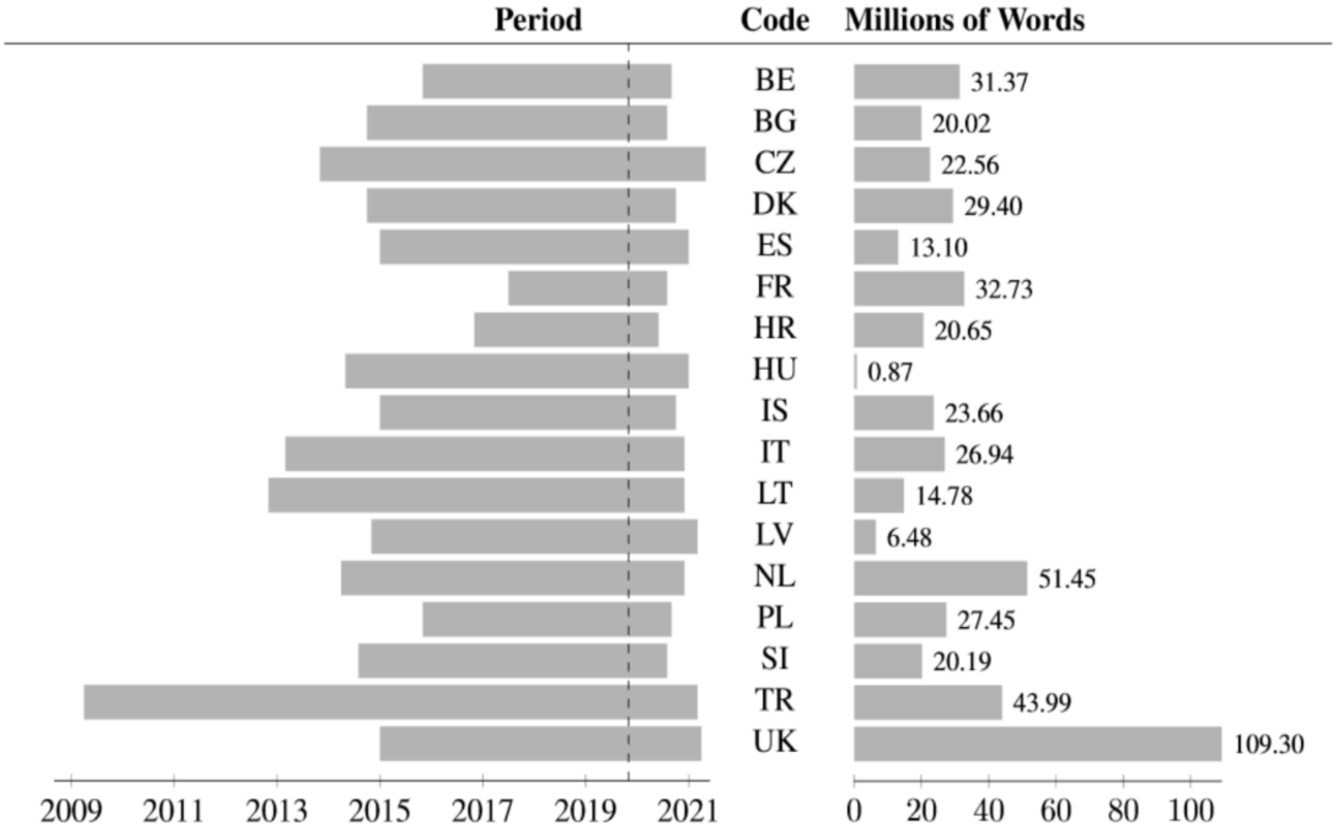
Corpora - contents

- reference and COVID-19 sections
- rich metadata about the mandates, sessions, and speakers, their political party affiliations etc.
- linguistic annotations for NER and UD morphological features and syntax
- encoded to a common and very strict schema, so their format is not merely interchangeable but also interoperable

Corpora - use

- deposited in the CLARIN.SI repository
- released under CC BY licence
  - http://hdl.handle.net/11356/1432
  - http://hdl.handle.net/11356/1431
- mounted on NoSketch Engine and Kontext concordancers
- used in the Helsinki DH Hackathon 2021 and 2022
- actions needed:
  - *schema and metadata improvement*
  - *expanding timespan of the data*
  - *adding new parliaments*
  - *making corpora more comparable and useful to various interested communities*

Countries, time span and size of the corpora



Schema and Metadata Improvements

Leads: Tomaž Erjavec (IJS), Matyáš Kopp (UFAL)

- Encoding of ParlaMint I corpora followed the previously developed TEI-based Parla-CLARIN recommendations for encoding parliamentary corpora
- ParlaMint required much stricter encoding, so we started project by defining a RelaxNG schema for corpus validation, which was then refined during most of the lifetime of the ParlaMint I project
- ParlaMint II involves more than 20 partners that are all required to submit large and heavily annotated corpora with rich metadata
  - Parla-CLARIN schema has to be harmonized accordingly
  - Establish validation procedures that will result in correctly and consistently encoded ParlaMint II corpora
  - Provide more elaborate documentation; git management
  - Finding common ground in various political systems for adequately adding new metadata

Corpus Expansion

Lead: Tomaž Erjavec (IJS)

- Adding corpora for new parliaments will extend the ParlaMint scope in countries, languages and time to make it even more interesting for researchers.
- Adding new data to the existing parliamentary corpora will extend the covid period as well as will add information about other disruptive events, such as the Russian invasion in Ukraine
- Both new and updated corpora will be validated, converted to derived formats (plain text, metadata files, CoNLL-U, vertical files), mounted on the CLARIN.SI concordancers, and deposited in the CLARIN.SI repository

Partners

- Austria, Basque Country, Belgium, Bulgaria, Catalonia, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Greece, Hungary, Iceland, Italy, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Romania, Slovenia, Spain, Sweden, Turkey, United Kingdom

Corpus Enrichment 1/2

Lead: Nikola Ljubešić (IJS)

- Machine Translation:
  - of all non-English transcriptions into English at a sentence level and with an automatic post-editing module
  - opens the way for translingual comparative analyses among more than 20 national and regional parliaments
- Semantic tagging:
  - of the translated corpus with the UCREL Semantic Analysis System (resp. Paul Rayson), USAS
  - assign coarse-grained word senses
  - accuracy for English is 91 %

Corpus Enrichment 2/2

Lead: Nikola Ljubešić (IJS)

- Alignment with audio
  - proof-of-concept on three selected languages (Czech, Polish and Croatian)
  - min. 50 hours of high-quality audio alignment
  - code base
  - report of the used alignment procedure
- ParlaSpeech-HR:
  - the first freely-available dataset for training automatic-speech-recognition systems for Croatian
  - based on the ParlaMint I corpus and the available video recordings of the Croatian parliament
  - 1,816 hours (http://hdl.handle.net/11356/1494)

Engagement Activities

Leads: Darja Fišer (INZ), Çagrı Çöltekin (TUB)

- Tutorial
  - topic modelling
- Showcases
  - demonstrate the value of the ParlaMint corpora for SSH researchers
  - serve as an instrument for cross-disciplinary method and knowledge transfer
- Shared task
  - governing/opposition prediction
  - party affiliation
  - political ideologies
  - training data: ParlaMint I corpora, test data: ParlaMint II corpora

Beyond ParlaMint II

- More data
  - from the European Parliament
  - regional parliaments
  - national parliaments beyond Europe
  - historical data
- Link with other data sources
  - e.g. voting results, social media content, newspaper and TV news mentions
- Further extensions
  - multimodal corpora
  - gesture annotated corpora
  - live corpora produced and used as streamed and on the fly

Reference

The project and its results described in:

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darģis, Orsolya Ring, Ruben van Heusden, Maarten Marx & Darja Fišer. **The ParlaMint corpora of parliamentary proceedings.**
In: *Language Resources & Evaluation* (2022).
https://doi.org/10.1007/s10579-021-09574-0 (openly available)

Contact info

- Maciej Ogrodniczuk: maciej.ogrodniczuk@gmail.com
- Petya Osenova: petya@bultreebank.org