ParlaMint: Towards Comparable Parliamentary Corpora

Maciej Ogrodniczuk, Michał Rudolf
Institute of Computer Science, Polish Academy of Sciences

Petya Osenova, Kiril Simov
Institute for Information and Communication Technologies, Bulgarian Academy of Sciences

Tomaž Erjavec, Nikola Ljubešić
Jožef Stefan Institute, Ljubljana

Andrej Pančur
Institute of Contemporary History, Ljubljana

Our motivation

One of the most important aspects of processing of parliamentary data is its correspondence to the most recent events with global impact on human health, social life and economics such as the current COVID-19 pandemic. By comparing the data synchronically and diachronically within a cross-lingual context, the scientific and civil communities will be able to track pan-European discussion and be quickly updated on any emerging topic.

Our goals

Our main aims are:

- to compile a collection of parliamentary corpora in a number of languages and in a harmonized format, covering both the current data and older, reference data
- to process the corpora linguistically
- to index the data with popular concordancers so that interested parties can search and extract the relevant comparable information
- to show through use cases that our resources and technology serve the society's needs.

Expected outcome

Strategy and data availability: The project will establish a strategy for handling parliamentary data and processing in times of any emergency (COVID-19 is just a showcase). Thus, different reference corpora could be produced with parliamentary records from previous times with global crisis states, e.g. the great economic recession, periods of floods in Europe, the Ebola outbreak etc.

Standards development: The Parla-CLARIN encoding scheme will be further developed to cover more detailed and specific metadata and data across languages and parliaments. The corpora will serve as a baseline for further updates. Such uniform updates across the corpora would strongly support various methods of comparative research.

From showcasing to real applications: The availability of comparable multilingual parliamentary data (also made visible through concordancers and Parlameter) will boost research in the areas of digital humanities, linguistics, politology, sociology, psychology as well as in all the related branches of sciences.

What’s already there?

Phase 1 (July–September 2020):

The approach tested for 4 pilot languages: Bulgarian, Croatian, Slovene and Polish:

1. COVID-19 (November 2019 – July 2020) and reference (2015–2019) corpora of parliamentary proceedings were encoded in a specialisation of the Parla-CLARIN format, also featuring annotation of Named Entities and Universal Dependencies syntax
2. The corpora have been made available through NoSketch Engine: [https://www.clarin.si/noske/parlamint.cgi/](https://www.clarin.si/noske/parlamint.cgi/)
3. The corpora will be soon available at [http://hdl.handle.net/11356/1345](http://hdl.handle.net/11356/1345)

What’s coming next?

Phase 2 (October 2020 – May 2021):

1. The corpora will be extended with more languages (call for mini-grants will be issued end of October 2020).
2. Use cases in Political Sciences and Digital Humanities will be built based on the corpus data.

ParlaMint project is supported by CLARIN ERIC.

[www.clarin.eu/content/parlamint](http://www.clarin.eu/content/parlamint)

[maciej.ogrodniczuk@gmail.com, petya@bultreebank.org](mailto:maciej.ogrodniczuk@gmail.com, petya@bultreebank.org)