ParlaMint: Comparable Corpora of European Parliamentary Data

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Andrej Pančur, Nikola Ljubešić, Tommaso Agnoloni, Starkaður Barkarson, María Calzada Pérez, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Luciana D. de Macedo, Jesse de Does, Katrien Depuydt, Sascha Diwersy, Dorte Haltrup Hansen, Matyáš Kopp, Tomas Krilavičius, Giancarlo Luxardo, Maarten Marx, Vaidas Morkevičius, Costanza Navarretta, Paul Rayson, Orsolya Ring, Michał Rudolf, Kiril Simov, Steinþór Steingrímsson, István Üveges, Ruben van Heusden, Giulia Venturi

CLARIN annual conference 2021
September 27, 2021

The project

- A mini-project supported by CLARIN-ERIC
- Budget: 135,000 €
- 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 lock-downs and restrictive measures as well as on the consequences with respect to health, medical care systems, employment, etc. during pandemic times.

Implementation

- Phase 1 (July 2020 – Sep 2020):
  Compiled the corpora of Bulgarian, Croatian, Polish and Slovene parliamentary speeches:
  Multilingual comparable corpora of parliamentary debates ParlaMint 1.0. 2020. `http://hdl.handle.net/11356/1345`
- Phase 2 (Dec 2020 – May (June) 2021:
  - call for additional corpora, 13 respondents
  - version 2.0 released towards the end of the project
  - V2.0 used in the Helsinki Digital Humanities Hackathon
  - version 2.1 built on the experience of DHH & fixed some errors
  - **V2.1: 17 corpora (countries) with 16 languages and half a billion words**

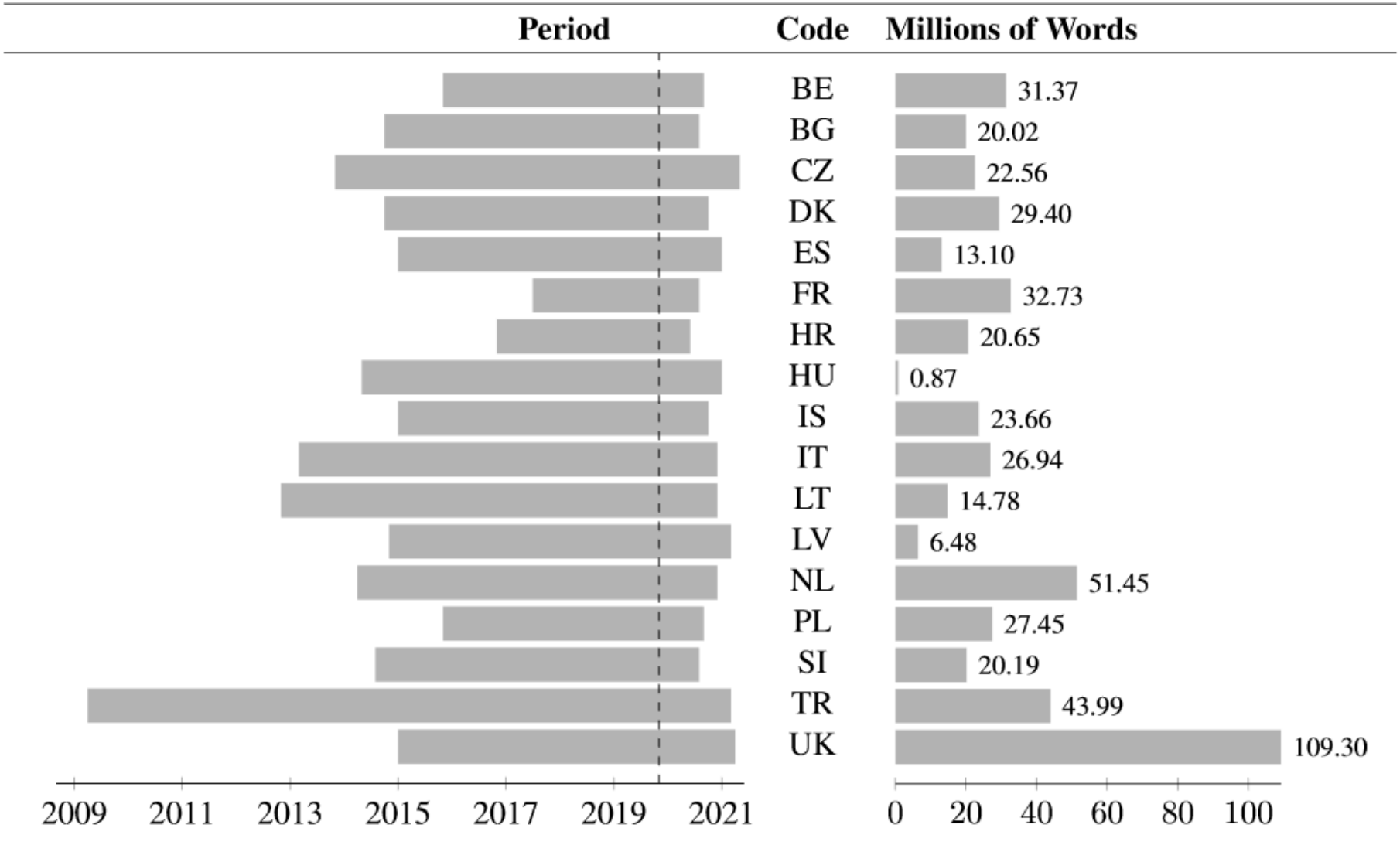
Results

- Downloadable sets of corpora (CC BY) @ CLARIN.SI:
  - Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. 2021. `http://hdl.handle.net/11356/1432`.
  - Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. 2021, `http://hdl.handle.net/11356/1431`
- Integrated with noSketch Engine and KonText
- Available through a dedicated instance of ParlaMeter, `https://parlamint.parlameter.org/`

GitHub

- Using Git was quite helpful for the project
  (but could've been even more so)
- **`https://github.com/clarin-eric/ParlaMint`**
  - XML schemas
  - samples of all corpora in ParlaMint (Parla-CLARIN/TEI) XML
  - also in the derived formats:
    plain text, TSV metadata files, CoNLL-U and vertical format
  - XSLT and Perl scripts for validation and conversion
  - some derived metadata information

Data overview



Information included

- Oppositions/coalitions, political parties
- Speakers with party memberships, MP status, gender
- Sessions with term/session number and date
- Speeches with speakers (chair, regular)
- Transcriber comments
- Linguistic annotation:
  - Tokens and sentences
  - Lemmas
  - UD PoS and morphological features
  - UD syntactic dependencies
  - Named entities (PER, LOC, ORG, MISC)

Conclusions

- Presented the ParlaMint project and corpora
- Further work in continuation project:
  - Better documentation, validation
  - Better Git(Hub) rules and control
  - More corpora
  - Extend current corpora in time and with metadata
  - MT all the corpora to English
  - Experiment with adding speech data
  - Using the corpora: DHH 2022, shared task, tutorial

ParlaMint: Comparable Corpora of European Parliamentary Data

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Andrej Pančur, Nikola Ljubešić, Tommaso Agnoloni, Starkaður Barkarson, María Calzada Pérez, Çağrı Çöltekin, Matthew Coole, Roberts Darģis, Luciana D. de Macedo, Jesse de Does, Katrien Depuydt, Sascha Diwersy, Dorte Haltrup Hansen, Matyáš Kopp, Tomas Krilavičius, Giancarlo Luxardo, Maarten Marx, Vaidas Morkevičius, Costanza Navarretta, Paul Rayson, Orsolya Ring, Michał Rudolf, Kiril Simov, Steinþór Steingrímsson, István Üveges, Ruben van Heusden, Giulia Venturi

CLARIN annual conference 2021
September 27, 2021