The ParlaMint corpora of parliamentary proceedings

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana

New Approaches to Editing Early Modern Parliamentary Records
Graz
April 6–8, 2022

Overview of the talk

Introduction

What is CLARIN?

- European research infrastructure for language resources and technologies
- Its goal is to support research communities from Humanities, Social Sciences and other language-related disciplines with:
  - language resources and technologies
  - expertise and knowledge transfer
- 22 member countries, each with at least one CLARIN centre

CLARIN work on Parliamentary corpora

- CLARIN Travelling Campus "Talk of Europe": three "Creative Camps" (2014–2015) used the proceedings of the European Parliament, curated as linked open data
- CLARIN-PLUS cross-disciplinary workshop Working with parliamentary records, Sofia 2017
- CLARIN Resource Families: Parliamentary corpora, 2018–2019
- First ParlaCLARIN workshop at LREC 2018
- CLARIN ParlaFormat workshop, Amersfoort, 2019
- Second ParlaCLARIN workshop at LREC 2020
- (Third ParlaCLARIN workshop at LREC 2022)
- **ParlaMint I**, 2020–2021
- **ParlaMint II**, 2022–2023

ParlaMint I

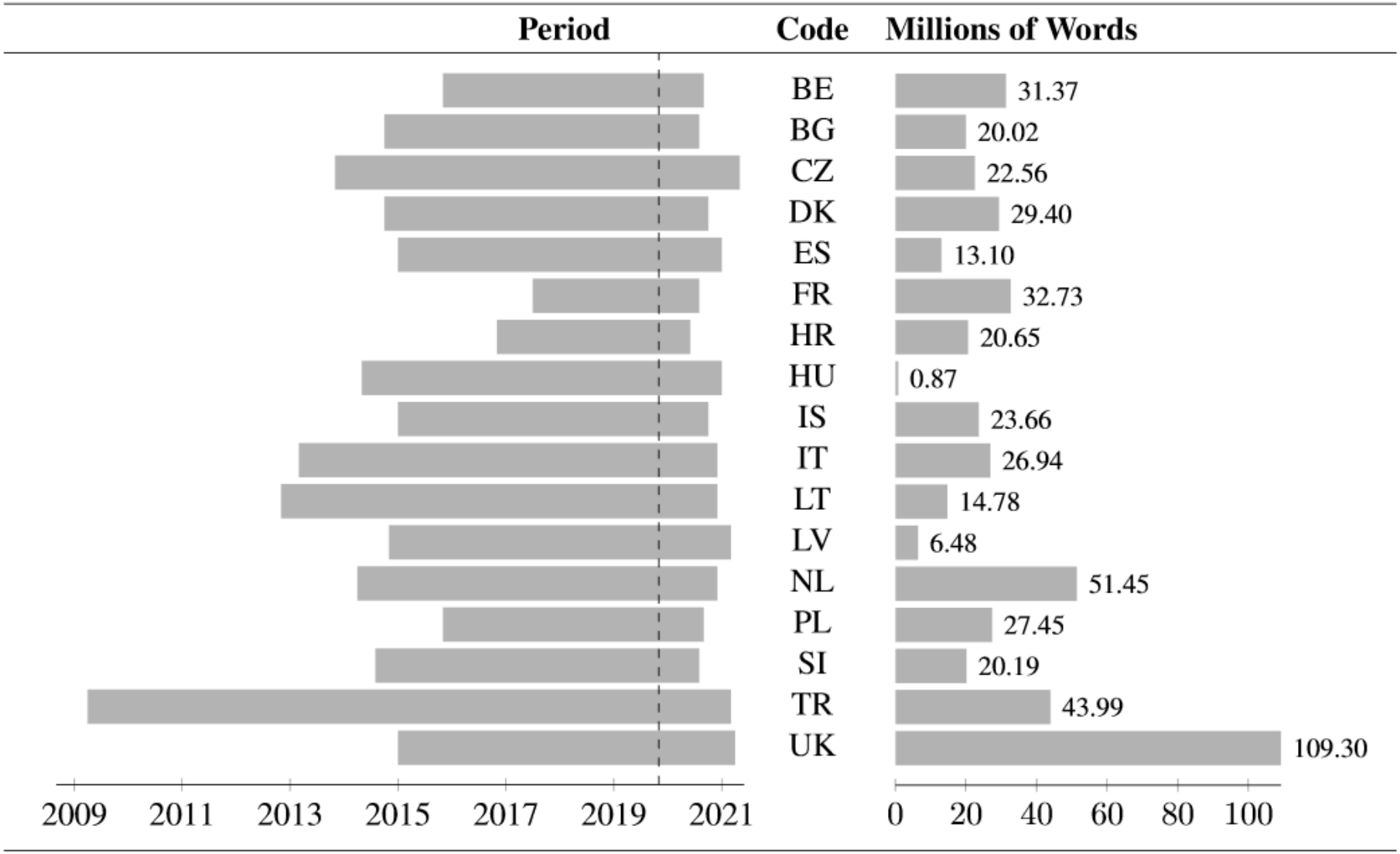
ParlaMint: the first CLARIN flagship project

- A mini-project supported by CLARIN-ERIC
- Budget: 135,000 €
- Duration: Jul 1 2020 – May 30 2021
- Motivation: **Parliamentary data** directly corresponds to events with global impact 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.

The ParlaMint corpora

- ParlaMint V2.1:
  - 17 corpora (countries)
  - 16 languages
  - 11,000 speakers
  - 500,000 speeches
  - 500,000,000 words
- Available for download under CC-BY:
  - 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
- Available on CLARIN.SI noSketch Engine and KonText
- V2.0 used in the Helsinki Digital Humanities Hackathon 2021

Countries, time span and size of the corpora



Information included

- Political parties, oppositions and coalitions
- Speakers with party memberships, MP status, gender
- Sessions with date and term/session/meeting number/title
- Speeches with marked speakers and their status (chair, regular)
- Text of the speeches
- Transcriber comments (gaps, interruptions, voting etc.)
- Linguistic annotation

The importance of encoding

- The idea of ParlaMint was that the corpora are encoded as uniformly as possible
- This would allow the corpora to be interoperable, so that e.g. they can be converted to other formats by the same scripts
- However:
  - the corpora had very different source encoding
  - they are differently structured, contain different information, and reflect different parliamentary traditions
  - each was corpus produced by a separate partner

The definition of a rich but constrained format and the possibility to validate the corpora was crucial.

Parla-CLARIN

- The "CLARIN ParlaFormat" workshop introduced a format for parliamentary corpora called "Parla-CLARIN" (Erjavec and Pančur, 2019)
- Parla-CLARIN is a simple TEI customisation with extensive annotation guidelines
- ParlaMint used the Parla-CLARIN encoding but further constrained it

ParlaMint validation and file formats

- Validation:
  1. bespoke Parla-CLARIN compatible RelaxNG schema
  2. XSLT script for content validation
  3. Errors in conversion to other formats
  4. Use of the data (in e.g. concordancers)
- Available formats:
  - Canonical Parla-CLARIN TEI XML encoding
  - Derived formats:
    - per-speech TSV tabular metadata (16 points)
    - plain text with speech ID
    - CoNLL-U format
    - vertical format with registry files (for concordancers)

GitHub

- Using Git was quite helpful for the project
  (but could've been even more so)
- https://github.com/clarin-eric/ParlaMint
  - ParlaMint annotation guidelines
  - XML schemas
  - samples of all corpora in ParlaMint XML and derived formats
  - XSLT and Perl scripts for validation and conversion
  - some derived metadata information

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.
Language Resources & Evaluation (2022).
https://doi.org/10.1007/s10579-021-09574-0

ParlaMint II

Continuation of the ParlaMint project

- February 2022 – May 2023
- Work packages:
  1. Upgrade schema, validation and workflow
  2. Extend existing corpora and add corpora for new parliaments
  3. Enhance the corpora with additional information
  4. Improve the use and usability of the corpora
- Lots to do...

WP1: Documentation, interoperability, metadata

- Upgrade Parla-CLARIN encoding guidelines in line with ParlaMint (done)
- Write ParlaMint encoding guidelines (done)
- Write ParlaMint TEI ODD schema specification (in progress)
- Implement validation of samples on GitHub CI (done)
- Identify bugs, reevaluate encoding choices, propose encodings for new types of (meta)data and linguistic annotation (in progress: so far 51 open issues)
- Add metadata on ministers and party orientation (starting)

WP Leads: Tomaž Erjavec and Matyáš Kopp

WP2: Corpus extension

Extend existing ParlaMint corpora to 2022-06 and fix errors.

Add new corpora:

- **Austria** (Tanja Wissik, Matej Durco, H. Pirker, K. Mörth)
- Basque Country (Mikel Iruskieta)
- Catalonia (Nuria Bel)
- Estonia (Kadri Vider, Neeme Kahusk, Martin Mölder)
- Finland (Eero Hyvönen, Jouni Tuominen)
- Greece (Maria Gavriilidou)
- Norway (M. Breder Birkenes, J. Arild Olsen, K. De Smedt)
- Portugal (Amália Mendes)
- Romania (Petru Rebeja, Madalina Chitez, Cornelia Ilie)
- Sweden (Fredrik Norén)

Also interested:

- Galicia (Elisa Fernandez Rey)
- Bosnia, Serbia (under discussion)

Lead: Tomaž Erjavec

WP3: Corpus enrichment

Add useful information to the corpora:

- Translate the corpora into English (OpenNMT)
- Semantic tagging (USAS: Paul Rayson, Lacaster University)
- Add recordings (proof of concept on Czech, Polish, Croatian)

Lead: Nikola Ljubešić

WP4: Engagement

- Helsinki **Digital Humanities** hackathon #DHH22, 11.–20.5.2022
- Shared task for **Natural Language Processing** (submitted to SemEval 2023)
- Tutorial for **Social Sciences and Humanities** scholars & students using ParlaMint and Orange
- Showcases demonstrating the use of ParlaMint to answer research questions from **SSH**

Leads: Darja Fišer, Çağrı Çöltekin

WP5: Coordination

- Management
- Dissemination
- External monitoring

Leads: Maciej Ogrodniczuk, Petya Osenova

Linguistic annotation

Linguistic annotation in ParlaMint I

- Tokens
- Sentences
- Lemmas
- (Per-language PoS tags, e.g. MULTEXT-East MSDs)
- Universal Dependencies PoS and morphological features
- UD syntactic dependencies
- Named entities (PER, LOC, ORG, MISC)
- Fine-grained named entities (Czech only)

An annotated sentence

From the ParlaMint-GB.ana corpus with text "1.":

```
<s xml:id="seg1.1">
  <w join="right"
     lemma="1"
     msd="UPosTag=NUM|NumType=Card"
     pos="LS"
     xml:id="seg1.1.1">1</w>
  <pc msd="UPosTag=PUNCT"
      pos="."
      xml:id="seg1.1.2">.</pc>
  <linkGrp targFunc="head argument" type="UD-SYN">
    <link ana="ud-syn:root"
          target="#seg1.1 #seg1.1.1"/>
    <link ana="ud-syn:punct"
          target="#seg1.1.1 #seg1.1.2"/>
  </linkGrp>
</s>
```

Annotation tools used

Each partner was free to choose which annotation tools to use.

- Morphology and syntax:
  - Core NLP: GB
  - UD Pipe: CZ, DK, IS, TR
  - Stanza: ES, FR, IT
  - CLASSLA: BG, HR, SI
  - Spacy: LT
  - UDify: NL
  - (Also) local annotation tools: BE, DK, IS, LV, NL, PL, TR
- NER: (almost) each one their own

The missing link

- New Approaches to Editing Early Modern Parliamentary Records
- Language – in particular words – are in historical texts different from the contemporary standard:
  - difficult to search in corpora (variability in spelling)
  - low performance of standard linguistic annotation tools (trained on contemporary language)
- Solution: modernisation of word-forms
- Normalisation:
  - modernisation of historical language
  - standardisation of user generated content or dialect transcriptions

CSMTiser

- CSMTiser = Character-based Statistical Machine Translation
- https://github.com/clarinsi/csmtiser
- Translation is orthographical!
- Needs training data:
  - lexicon of historical : contemporary word pairs
    ⇒ translation model
  - large corpus of contemporary language
    ⇒ target language model
- References:
  - Ljubešić, Nikola, Katja Zupan, Darja Fišer and Tomaž Erjavec. 2016. Normalising Slovene data: historical texts vs. user-generated content. KONVENS.
  - Scherrer, Yves, and Nikola Ljubešić. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. KONVENS.

Conclusions

Conclusions

- Presented the completed project ParlaMint I
- and the recently started ParlaMint II
- Some discussion on the linguistic annotation tools we used
- and on word normalisation

The ParlaMint corpora of parliamentary proceedings

Tomaž Erjavec
Department of Knowledge Technologies
Jožef Stefan Institute
Ljubljana

New Approaches to Editing Early Modern Parliamentary Records
Graz
April 6–8, 2022