

LREC 2020 Workshop  
Language Resources and Evaluation Conference  
11–16 May 2020

**Creating, Using and Linking of Parliamentary  
Corpora with Other Types of Political Discourse  
( ParlaCLARIN II)**

**PROCEEDINGS**

Editors: Darja Fišer, Maria Eskevich, Franciska de Jong

**Proceedings of the LREC 2020 Workshop on  
Creating, Using and Linking of Parliamentary Corpora  
with Other Types of Political Discourse  
(ParlaCLARIN II)**

Edited by: Darja Fišer, Maria Eskevich, Franciska de Jong

Acknowledgement: Organisation of the workshop is supported by CLARIN ERIC.  
Workshop web site: <https://www.clarin.eu/ParlaCLARIN-II>

**ISBN: 979-10-95546-47-4**

**EAN: 9791095546474**

**For more information:**

European Language Resources Association (ELRA)

9 rue des Cordelières

75013, Paris

France

<http://www.elra.info>

Email: [lrec@elda.org](mailto:lrec@elda.org)

© European Language Resources Association (ELRA)

These workshop proceedings are licensed under a Creative Commons  
Attribution-NonCommercial 4.0 International License

## Introduction

Parliamentary data is a major source of scholarly and socially relevant content. In line with the trend towards open access to digitised sources of materials from the public domain, parliamentary records are available in ever larger quantities and the number of languages for which such records are made public is also increasing. The materials are accompanied by rich metadata, and in addition there is the characteristic that the original spoken word sources, traditionally archived in transcribed form only, are now increasingly released also in audio and video formats. In linguistically and culturally diverse regions such as Europe there is harmonization of the data curation practices can advance the field significantly. As a result there is immense potential for developing models for collaborative work, comparative perspectives and multidisciplinary research, and we envision that the ParlaCLARIN II workshop will contribute to the further articulation of this agenda.

An inspiring and highly successful first edition of the ParlaCLARIN scientific workshop<sup>1</sup> was held at LREC 2018 and a follow-up developmental workshop was organized by CLARIN ERIC in 2019 under the name ParlaFormat<sup>2</sup>. These events led to a comprehensive overview<sup>3</sup> of a multitude of existing parliamentary resources worldwide as well as tangible first steps towards better harmonization, interoperability and comparability of the resources and tools relevant for the study of parliamentary discussions and decisions.

The second ParlaCLARIN workshop aims to broaden the scope of research questions that is enabled thanks to proposed data harmonisation approach, as well as to continue the effort of building a multidisciplinary community around parliamentary data by bringing together developers, curators and researchers of regional, national and international parliamentary debates that are suitable for research in disciplines in the Humanities and Social Sciences. We invited unpublished original work focusing on the compilation, annotation, visualisation and utilisation of parliamentary records as well as linking or comparing parliamentary records with other datasets of political discourse such as party manifestos, political speeches, political campaign debates, social media posts, etc. Apart from dissemination of the results, the workshop also addresses the identified obstacles, discusses open issues and coordinates future efforts in this increasingly trans-national and cross-disciplinary community.

The accepted papers address the following topics:

- Creation and annotation of parliamentary data in textual and spoken format
- Enrichment of parliamentary data with semantic and name entity tagging
- Querying and visualisation of parliamentary data
- Text mining over parliamentary and other political language data
- Comparative studies of parliamentary corpora
- Diachronic studies based on parliamentary corpora

The workshop programme is composed of a keynote talk by Pola Lehmann and Bernhard Weßels and 13 peer-reviewed papers by 30 authors from 10 countries.

---

<sup>1</sup><https://www.clarin.eu/ParlaCLARIN>

<sup>2</sup><https://www.clarin.eu/event/2019/parlaformat-workshop>

<sup>3</sup><https://www.clarin.eu/resource-families/parliamentary-corpora>

We would like to thank the reviewers for their careful and constructive reviews which have contributed to the quality of the event.

Due to the COVID-19 pandemic, the main conference is postponed, for this reason, there will be only a virtual ParlaCLARIN II meeting in May 2020.

D. Fišer, M. Eskevich, F. de Jong

May 2020

**Organizers:**

Darja Fišer, University of Ljubljana and Jožef Stefan Institute, Slovenia  
Franciska de Jong, CLARIN ERIC, The Netherlands  
Maria Eskevich, CLARIN ERIC, The Netherlands

**Program Committee:**

Kaspar Beelen, The Alan Turing Institute, UK  
Andreas Blätte, The University of Duisburg-Essen, Germany  
Tomaž Erjavec, Jožef Stefan Institute, Slovenia  
Francesca Frontini, Université Paul Valéry - Montpellier, France  
Maria Gavriilidou, ILSP/Athena RC, Greece  
Henk van den Heuvel, Radboud University, The Netherlands  
Klaus Illmayer, Austrian Academy of Sciences, Austria  
Bente Maegaard, CLARIN ERIC, The Netherlands  
Monica Monachini, National Research Council of Italy, Italy  
Laura Morales, Sciences Po, France  
Jan Odijk, Utrecht University, The Netherlands  
Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, Poland  
Petya Osenova, IICT-BAS and Sofia University "St. Kl. Ohridski", Bulgaria  
Stelios Piperidis, ILSP/Athena RC, Greece  
Maria Pontiki, ILSP/Athena RC, Greece  
Simone Paolo Ponzetto, University of Mannheim, Germany  
Sara Tonelli, Fondazione Bruno Kessler, Italy  
Tamás Váradi, Hungarian Academy of Sciences, Hungary  
Tanja Wissik, Austrian Academy of Sciences, Austria

**Invited Speaker:**

Pola Lehmann and Bernhard Weßels, Berlin Social Science Center (WZB), Germany

# Workshop Program

## Creation of parliamentary corpora

### *New Developments in the Polish Parliamentary Corpus*

Maciej Ogródniczuk and Bartłomiej Nitoń ..... 1

### *Anföranden: Annotated and Augmented Parliamentary Debates from Sweden*

Stian Rødven Eide ..... 5

### *IGC-Parl: Icelandic Corpus of Parliamentary Proceedings*

Steinþór Steingrímsson, Starkaður Barkarson and Gunnar Thor Örnólfsson ..... 11

### *Compiling Czech Parliamentary Stenographic Protocols into a Corpus*

Barbora Hladka, Matyáš Kopp and Pavel Straňák ..... 18

### *Unfinished Business: Construction and Maintenance of a Semantically Tagged Historical Parliamentary Corpus, UK Hansard from 1803 to the present day*

Matthew Coole, Paul Rayson and John Mariani ..... 23

### *The siParl corpus of Slovene parliamentary proceedings*

Andrej Pancur and Tomaž Erjavec ..... 28

## Tools for parliamentary corpora

### *Who mentions whom? Recognizing political actors in proceedings*

Lennart Kerkvliet, Jaap Kamps and Maarten Marx ..... 35

### *Challenges of Applying Automatic Speech Recognition for Transcribing EU Parliament Committee Meetings: A Pilot Study*

Hugo de Vos and Suzan Verberne ..... 40

### *Parsing Icelandic Alþingi Transcripts: Parliamentary Speeches as a Genre*

Kristján Rúnarsson and Einar Freyr Sigurðsson ..... 44

## **Investigations of parliamentary corpora**

### *Identifying Parties in Manifestos and Parliament Speeches*

Costanza Navarretta and Dorte Haltrup Hansen ..... 51

### *Comparing Lexical Usage in Political Discourse across Diachronic Corpora*

Klaus Hofmann, Anna Marakasova, Andreas Baumann, Julia Neidhardt and Tanja Wissik ..... 58

### *The Europeanization of Parliamentary Debates on Migration in Austria, France, Germany, and the Netherlands*

Andreas Blaette, Simon Gehlhar and Christoph Leonhardt ..... 66

### *Querying a large annotated corpus of parliamentary debates*

Sascha Diwersy and Giancarlo Luxardo ..... 75