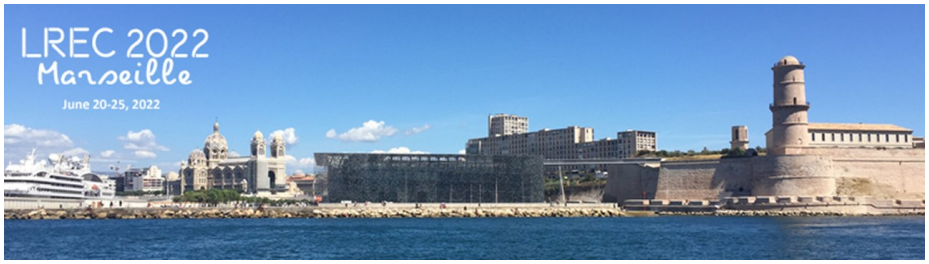

LREC 2022
Language Resources and Evaluation Conference
20-25 June 2022

ParlaCLARIN III

Workshop on Creating, Enriching and Using Parliamentary Corpora

PROCEEDINGS

Editors: Darja Fišer, Maria Eskevich, Jakob Lenardič,
Franciska de Jong

Proceedings of the LREC 2022 ParlaCLARIN III Workshop on Creating, Enriching and Using Parliamentary Corpora

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

ISBN: 979-10-95546-85-6
EAN: 9791095546856

Acknowledgements: Organisation of the workshop is supported by CLARIN ERIC.
https://www.clarin.eu/ParlaCLARIN-III

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Introduction

Parliamentary data is an important source of scholarly and socially relevant content, serving as a verified communication channel between the elected political representatives and members of the society. The development of accessible, comprehensive and well-annotated parliamentary corpora is therefore crucial for the information society, as such corpora help scientists and investigative journalists to ascertain the accuracy of socio-politically relevant information, and to inform the citizens about the trends and insights on the basis of such data explorations. Research-wise, parliamentary corpora are a quintessential resource for a number of disciplines in digital humanities and social sciences, such as political science, sociology, history, and (socio)linguistics.

The distinguishing characteristic of parliamentary data is that it is spoken language produced in controlled circumstances. Such data has traditionally been transcribed in a formal way but is now also increasingly released in the original audio and video formats, which encourages resource and software development and provides research opportunities related to structuring, synchronisation, visualisation, querying and analysis of parliamentary corpora. Therefore, a harmonised approach to data curation practises for this type of data can support the advancement of the field significantly. One of the ways in which the research community is supported in this line of work is through the conversion of existing corpora and further development of new cross-national parliamentary corpora into a highly comparable, harmonised set of multilingual resources. These allow researchers to share comparative perspectives and to perform multidisciplinary research on parliamentary data. We envision that the ParlaCLARIN III workshop, as a venue for knowledge and experience exchange on the topic, will contribute to the development and growth of the field of digital parliamentary science.

An inspiring and highly successful first edition of the ParlaCLARIN scientific workshop[1] was held at LREC 2018. A follow-up developmental workshop was organised by CLARIN ERIC in 2019 under the name ParlaFormat[2], while the second ParlaCLARIN workshop was held at LREC 2020.[3] These events led to a comprehensive overview[4] of a multitude of existing parliamentary resources worldwide as well as tangible first steps towards better harmonisation, interoperability and comparability of the resources and tools relevant for the study of parliamentary discussions and decisions.

This third ParlaCLARIN workshop is a continuation of the 2018 and 2020 editions. On the one hand, it continues to bring together developers, curators and researchers of regional, national and international parliamentary debates from across diverse disciplines in the Humanities and Social Sciences. On the other hand, we envisage the appearance of new discussion threads, tasks, and challenges that are partially inspired by or related to the new data releases such as ParlaMint [5] and data formats such as Parla-CLARIN.[6]

The Call for Papers has invited original, overview and position papers with the focus on one of the following topics:

- Compilation, annotation, visualisation and utilisation of parliamentary records;
- Harmonisation of existing multilingual parliamentary resources, containing either synchronic or diachronic data or both;
- Linking or comparing of parliamentary records with other sources of structured knowledge, such as formal ontologies and LOD datasets (in particular for the description of speakers, political parties, etc.).

[1] https://www.clarin.eu/ParlaCLARIN
[2] https://www.clarin.eu/event/2019/parlaformat-workshop
[3] https://www.clarin.eu/ParlaCLARIN-II
[4] https://www.clarin.eu/resource-families/parliamentary-corpora
[5] https://www.clarin.eu/parlamint
[6] https://github.com/clarin-eric/parla-clarin

In 2022 the following special themes were also brought for discussion at the workshop:

- Machine translation of parliamentary proceedings and research using machine translated parliamentary data;
- Semantic tagging of parliamentary proceedings and research using semantically tagged parliamentary data;
- Digital Humanities and Social Sciences research into parliamentary proceedings.

The workshop programme is composed of a keynote talk by Luke Blaxill from the University of Oxford and 18 peer-reviewed papers by 66 authors from 15 countries (the 5 most represented: Germany (10), Italy (10), Slovenia (8), Austria (6) Spain (6)). Two papers report on the work that was carried out by the co-authors representing the institutions in more than one country, and one group of authors represent Canadian studies.

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

The ParlaCLARIN III workshop was held in person with the a possibility of hybrid attendance in Marseille (France), as part of the 13th edition of the Language Resources and Evaluation Conference (LREC2022).

D. Fišer, M. Eskevich, J. Lenardič , F. de Jong June 2022

Organizers

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

Program Committee:

Ahlame Bedgouri, Faculty of Sciences and Technology of Fez, University of Sidi Mohamed Ben Abdellah, Morocco
Çağrı Çöltekin, University of Tübingen, Germany
Jesse de Does, Dutch Language Institute, The Netherlands
Tomaž Erjavec, Jožef Stefan Institute, Slovenia
Francesca Frontini, Istituto di Linguistica Computazionale "A. Zampolli", CNR Pisa, Italy
Maria Gavriilidou, ILSP/Athena RC, Greece
Barbora Hladká, Charles University, Czechia
Haidee Kotze, Utrecht University, The Netherlands
Nikola Ljubešić, Jožef Stefan Institute, Slovenia
Bente Maegaard, CST, Department of Nordic Languages and Linguistics, University of Copenhagen, Denmark
Maarten Marx, University of Amsterdam, The Netherlands
Stefano Menini, Fondazione Bruno Kessler, Trento, Italy
Robert Muthuri, Anjarwalla & Khanna LLP, Kenya
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
Simone Paolo Ponzetto, Mannheim University, Germany
Paul Rayson, Lancaster University, United Kingdom
Sara Tonelli, Fondazione Bruno Kessler, Italy
Daniela Trotta, University of Salerno, Italy

Invited Speaker:

Luke Blaxill, University of Oxford, United Kingdom

Table of Contents

Conference Program

20 June 2022

9:15–9:30 Welcome and Introduction

9:30–10:30 Session 1: Corpus Creation 1

ParlaMint II: The Show Must Go On
Maciej Ogrodniczuk, Petya Osenova, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Çağrı Çöltekin, Matyáš Kopp and Meden Katja

How GermaParl Evolves: Improving Data Quality by Reproducible Corpus Preparation and User Involvement
Andreas Blaette, Julia Rakers and Christoph Leonhardt

Between History and Natural Language Processing: Study, Enrichment and Online Publication of French Parliamentary Debates of the Early Third Republic (1881-1899)
Marie Puren, Aurélien Pellet, Nicolas Bourgeois, Pierre Vernus and Fanny Lebreton

A French Corpus of Québec's Parliamentary Debates
Pierre André Ménard and Desislava Aleksandrova

11:00–12:00 Keynote

Parliamentary Corpora and Research in Political Science and Political History
Luke Blaxill

12:00–13:00 Session 2: Corpus Enhancement

Error Correction Environment for the Polish Parliamentary Corpus
Maciej Ogrodniczuk, Michał Rudolf, Beata Wójtowicz and Sonia Janicka

Clustering Similar Amendments at the Italian Senate
Tommaso Agnoloni, Carlo Marchetti, Roberto Battistoni and Giuseppe Briotti

Entity Linking in the ParlaMint Corpus
Ruben van Heusden, Maarten Marx and Jaap Kamps

Visualizing Parliamentary Speeches as Networks: the DYLEN Tool
Seung-bin Yim, Katharina Wünsche, Asil Cetin, Julia Neidhardt, Andreas Baumann and Tanja Wissik

20 June 2022 (continued)

14:00–15:15 Session 3: Corpus Analysis

Emotions Running High? A Synopsis of the state of Turkish Politics through the ParlaMint Corpus
Gül M. Kurtoğlu Eskişar and Çağrı Çöltekin

Immigration in the Manifestos and Parliament Speeches of Danish Left and Right Wing Parties between 2009 and 2020
Costanza Navarretta, Dorte Haltrup Hansen and Bart Jongejan

Parliamentary Discourse Research in Sociology: Literature Review
Jure Skubic and Darja Fišer

FrameASt: A Framework for Second-level Agenda Setting in Parliamentary Debates through the Lense of Comparative Agenda Topics
Christopher Klamm, Ines Rehbein and Simone Paolo Ponzetto

Comparing Formulaic Language in Human and Machine Translation: Insight from a Parliamentary Corpus
Yves Bestgen

15:15–16:00 Panel

16:30–17:45 Session 4: Corpus Creation 2

Adding the Basque Parliament Corpus to ParlaMint Project
Jon Alkorta and Mikel Iruskieta Quintian

ParlaSpeech-HR - a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus
Nikola Ljubešić, Danijel Koržinek, Peter Rupnik and Ivo-Pavao Jazbec

Making Italian Parliamentary Records Machine-Actionable: the Construction of the ParlaMint-IT corpus
Tommaso Agnoloni, Roberto Bartolini, Francesca Frontini, Simonetta Montemagni, Carlo Marchetti, Valeria Quochi, Manuela Ruisi and Giulia Venturi

ParlamentParla: A Speech Corpus of Catalan Parliamentary Sessions
Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos and Marta Villegas

ParlaMint-RO: Chamber of the Eternal Future
Petru Rebeja, Mădălina Chitez, Roxana Rogobete, Andreea Dincă and Loredana Bercuci

17:45–18:00 Pitches of relevant initiatives in the field