LREC-COLING 2024

ParlaCLARIN IV
Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora

Proceedings

Editors
Darja Fišer, Maria Eskevich, David Bordon

May 20, 2024
Torino, Italia

Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024): ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora

ISBN 978-2-493814-24-1
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

ParlaCLARIN IV @ LREC-COLING2024: 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 transcribed with speech-to-text software as well as 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 practices 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 IV 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.

This fourth ParlaCLARIN workshop is a continuation of the 2018[1], 2020[2] and 2022[3] editions held at the respective LREC conferences, see references below. 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[4] and data formats such as ParlaCLARIN[5] .

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 historical or contemporary parliamentary written or audio records
- Harmonisation of existing multilingual parliamentary resources, containing either synchronic or diachronic data or both
- Linking or comparing of parliamentary records with other datasets of political discourse such as party manifestos, political speeches, political campaign debates, and social media posts, and to 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/ParlaCLARIN-II
[3] https://www.clarin.eu/ParlaCLARIN-III
[4] https://www.clarin.eu/parlamint
[5] https://github.com/clarin-eric/parla-clarin

In 2024 the following special themes were also brought for discussion at the workshop:

- Enrichment of parliamentary proceedings (with e.g. sentiment annotation, political profiling of speakers etc.) and research using such data
- Machine translation of parliamentary proceedings and research using such data
- Argument mining of parliamentary debates

The workshop programme is composed of a keynote talk by Ines Rehbein from the Universität Mannheim and 24 peer-reviewed papers (of which 8 are presented as posters and 5 as demos) by 69 authors from 15 countries (the three most represented: Germany (5), Slovenia (4) and Czech Republic (3)). Two papers report on the work that was carried out by the co-authors representing the institutions in more than one country.

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

The ParlaCLARIN IV workshop was held in person with the a possibility of hybrid attendance in Turin (Italy), as part of the 2024 Joint International Confrence on Computational Linguistics, Language Resources and Evaluation (LREC-COLING2024).

D. Fišer, M. Eskevich, D. Bordon May 2024

Organizing Committee

- Darja Fišer, Institute of Contemporary History & CLARIN
- Maria Eskevich, Huygens Institute, KNAW
- David Bordon, University of Ljubljana

Program Committee

- Kaspar Beelen, The Alan Turing Institute, GB
- Siddharth Bhargava, Fondazione Bruno Kessler, IT
- Andreas Blaette, University of Duisburg-Essen, DE
- Hajo Boomgaarden, University of Vienna, AT
- Robert Borges, Uppsala University, SE
- Çağrı Çöltekin, University of Tübingen, DE
- Tomaž Erjavec, Dept. of Knowledge Technologies, Jožef Stefan Institute, SI
- Francesca Frontini, Istituto di Linguistica Computazionale "A. Zampolli" - ILC Consiglio Nazionale delle Ricerche - CNR, IT
- Maria Gavriilidou, ILSP / Athena RC, GR
- Turo Hiltunen, University of Helsinki, FI
- Pasi Ihalainen, University of Jyväskylä, FI
- Tatsuya Kawahara, Kyoto University, JP
- Haidee Kotze, Utrecht University, NL
- Anna Kryvenko, NISS (Ukraine); INZ (Slovenia), UA
- Cristina Lastres-López, University of Seville, ES
- Bente Maegaard, University of Copenhagen, DK
- Christian Mair, University of Freiburg, DE
- Maarten Marx, University of Amsterdam, NL
- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" - CNR, IT
- Jan Odijk, Utrecht University, NL
- Maciej Ogrodniczuk, Institute of Computer Science, Polish Academy of Sciences, PL
- Petya Osenova, Sofia University "St. Kl. Ohridski" and IICT-BAS, BG
- Stelios Piperidis, Athena RC/ILSP, GR
- Maria Pontiki, Institute for Language and Speech Processing (ILSP), Athena R.C., Greece, GR

- Simone Paolo Ponzetto, University of Mannheim, DE
- Valeria Quochi, Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale "A. Zampolli", IT
- Hugo Sanjurjo-González, University of Deusto, ES
- Sara Tonelli, FBK, IT
- Turo Vartiainen, University of Helsinki, FI
- Tanja Wissik, Austrian Academy of Sciences, AT

Invited Speaker

- Ines Rehbein, University of Mannheim Data and Web Science Group, DE

Table of Contents

ParlaCLARIN IV Workshop Program

20 May 2024

9:00–9:10 **Welcome and Introduction**

9:10–10:30 **ParlaMint**

9:10–9:30 *Parliamentary Discourse Research in Political Science: Literature Review*
Jure Skubic and Darja Fišer

9:30–9:50 *Compiling and Exploring a Portuguese Parliamentary Corpus: ParlaMint-PT*
José Aires, Aida Cardoso, Rui Pereira and Amalia Mendes

9:50–10:10 *Gender, Speech, and Representation in the Galician Parliament: An Analysis Based on the ParlaMint-ES-GA Dataset*
Adina I. Vladu, Elisa Fernández Rei, Carmen Magariños and Noelia García Díaz

10:10–10:30 *Bulgarian ParlaMint 4.0 corpus as a testset for Part-of-speech tagging and Named Entity Recognition*
Petya Osenova and Kiril Simov

11:00–12:00 **Keynote**

11:00–12:00 *Resources and Methods for Analysing Political Rhetoric and Framing in Parliamentary Debates*
Ines Rehbein

20 May 2024 (continued)

12:00–12:40 Creation of Parliamentary Language Resources

12:00–12:20 *PTPARL-V: Portuguese Parliamentary Debates for Voting Behaviour Study*
Afonso Sousa and Henrique Lopes Cardoso

12:20–12:40 *Polish Round Table Corpus*
Maciej Ogrodniczuk, Ryszard Tuora and Beata Wójtowicz

14:00–15:00 Analysis of Parliamentary Discourse

14:00–14:20 *Investigating Multilinguality in the Plenary Sessions of the Parliament of Finland with Automatic Language Identification*
Tommi Jauhiainen, Jussi Piitulainen, Erik Axelson, Ute Dieckmann, Mietta Lennes, Jyrki Niemi, Jack Rueter and Krister Lindén

14:20–14:40 *Exploring Word Formation Trends in Written, Spoken, Translated and Interpreted European Parliament Data – A Case Study on Initialisms in English and German*
Katrin Menzel

14:40–15:00 *Quantitative Analysis of Editing in Transcription Process in Japanese and European Parliaments and its Diachronic Changes*
Tatsuya Kawahara

15:00–15:40 Language Technology for Parliamentary Discourse

15:00–15:20 *Automated Emotion Annotation of Finnish Parliamentary Speeches Using GPT-4*
Otto Tarkka, Jaakko Koljonen, Markus Korhonen, Juuso Laine, Kristian Martiskainen, Kimmo Elo and Veronika Laippala

15:20–15:40 *Making Parliamentary Debates More Accessible: Aligning Video Recordings with Text Proceedings in Open Parliament TV*
Olivier Aubert and Joscha Jäger

20 May 2024 (continued)

15:40–16:00 Poster pitches

16:30–17:45 Poster session

16:30–17:45 *Russia and Ukraine through the Eyes of ParlaMint 4.0: A Collocational CADS Profile of Spanish and British Parliamentary Discourses*
Maria Calzada Perez

16:30–17:45 *Multilingual Power and Ideology identification in the Parliament: a reference dataset and simple baselines*
Çağrı Çöltekin, Matyáš Kopp, Meden Katja, Vaidas Morkevicius, Nikola Ljubešić and Tomaž Erjavec

16:30–17:45 *IMPAQTS: a multimodal corpus of parliamentary and other political speeches in Italy (1946-2023), annotated with implicit strategies*
Federica Cominetti, Lorenzo Gregori, Edoardo Lombardi Vallauri and Alessandro Panunzi

16:30–17:45 *ParlaMint Ngram viewer: Multilingual Comparative Diachronic Search Across 26 Parliaments*
Asher de Jong, Taja Kuzman, Maik Larooij and Maarten Marx

16:30–17:45 *Investigating Political Ideologies through the Greek ParlaMint corpus*
Maria Gavriilidou, Dimitris Gkoumas, Stelios Piperidis and Prokopis Prokopidis

16:30–17:45 *ParlaMint in TEITOK*
Maarten Janssen and Matyáš Kopp

16:30–17:45 *Historical Parliamentary Corpora Viewer*
Alenka Kavčič, Martin Stojanoski and Matija Marolt

16:30–17:45 *The dbpedia R Package: An Integrated Workflow for Entity Linking (for ParlaMint Corpora)*
Christoph Leonhardt and Andreas Blaette

16:30–17:45 *Video Retrieval System Using Automatic Speech Recognition for the Japanese Diet*
Mikitaka Masuyama, Tatsuya Kawahara and Kenjiro Matsuda

16:30–17:45 *One Year of Continuous and Automatic Data Gathering from Parliaments of European Union Member States*
Ota Mikušek

20 May 2024 (continued)

16:30–17:45 *Government and Opposition in Danish Parliamentary Debates*
Costanza Navarretta and Dorte Haltrup Hansen

16:30–17:45 *A new Resource and Baselines for Opinion Role Labelling in German Parliamentary Debates*
Ines Rehbein and Simone Paolo Ponzetto

16:30–17:45 *ParlaMint Widened: a European Dataset of Freedom of Information Act Documents (Position Paper)*
Gerda Viira, Maarten Marx and Maik Larooij

17:45–18:00 Closing Remarks