



LREC 2026

**ParlaCLARIN V:
Workshop on Interoperability, Multilinguality,
and Multimodality in Parliamentary Corpora**

Workshop Proceedings

**Editors
Maria Eskevich, Vincent Vandeghinste, David Bordon**

16 May 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-
NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-81-4

ParlaCLARIN V @ LREC2026: Preface

April 22, 2026

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 fifth ParlaCLARIN workshop is a continuation of the 2018¹, 2020², 2022³, and 2024⁴ 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 ParlaMint⁵ and data formats such as ParlaCLARIN⁶.

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

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/ParlaCLARIN-IV>

⁵<https://www.clarin.eu/parlamint>

⁶<https://github.com/clarin-eric/parla-clarin>

- 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.)

In 2026 the following special themes were also brought for discussion at the workshop:

- Multimodal corpora, containing video, speech, speech transcripts, official proceeding and research using such data
- Multilingual corpora, possibly containing live interpretation into spoken or signed languages and research using such data
- Position papers on the status and the future of the field of parliamentary corpus collection and research

The half-day workshop programme is composed of a keynote talk by Nikola Ljubešić from Jožef Stefan Institute in Ljubljana, Slovenia and 8 peer-reviewed papers by 19 authors from 6 countries.

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

The ParlaCLARIN V workshop was held in person with the a possibility of hybrid attendance in Palma de Mallorca (Spain), as part of the 2026 International Conference on Language Resources and Evaluation (LREC2026).

M. Eskevich, V. Vandeghinste, D. Bordon

May 2026

Organizing Committee

- Maria Eskevich, Huygens Institute, KNAW
- Vincent Vandeghinste, Dutch Language Institute & KU Leuven
- David Bordon, University of Ljubljana

Programme Committee

- Robert Borges, Uppsala University, SE
- Çağrı Çöltekin, University of Tübingen, DE
- Tomaž Erjavec, Jožef Stefan Institute, SI
- Maria Gavrilidou, Athena Research Center, GR
- Francesca Frontini, CNR-ILC, IT
- Pasi Ihalainen, University of Jyväskylä, FI
- Cristina Lastres-López, Universidad de Sevilla, ES
- Bente Maegaard, University of Copenhagen, DK
- Christian Mair, University of Freiburg, DE
- Maarten Marx, University of Amsterdam, NL
- Michal Mochtak, Radboud University, NL
- Jan Odijk, Utrecht University, NL
- Maciej Ogrodniczuk, Polish Academy of Sciences, PL
- Petya Osenova, Sofia University, BG
- Paul Rayson, Lancaster University, UK
- Hugo Sanjurjo-González, University of Deusto, ES
- Sara Tonelli, Fondazione Bruno Kessler, IT

Table of Contents

<i>From Concordance to Inference: ParlaCAP Helps ParlaMint Escape the Linguistics Lab</i> Nikola Ljubešić	1
<i>Quantifying Code-Switching in a Ukrainian Parliamentary Dataset 1990-2021</i> Olha Kanishcheva and Maria Shvedova	2
<i>Ours and Yours: A Discourse Analysis of Political Identity Markers in Slovenian Parliamentary Discourse</i> Meden Katja	13
<i>Representations of Europe and the European Union in Parliamentary Discourse from a Corpus-Assisted Perspective</i> Anna Kryvenko	22
<i>Towards ParlaMint-DE: Improving the Interoperability of the GermaParl Corpus of Plenary Protocols of the German Bundestag</i> Christoph Leonhardt and Andreas Blätte	31
<i>From Transcripts to Insights: A Digital Corpus and Interactive Speech Analysis Platform for Turkish Parliamentary Records</i> Basak Tepe, Irem Nur Yildirim, Onur Gungor and Susan Uskudarli	44
<i>Transcription and Recognition of Italian Parliamentary Speeches Using Vision-Language Models</i> Luigi Curini, Alfio Ferrara, Giovanni Pagano and Sergio Picascia	56
<i>Beyond OCR: Structural Segmentation and Speaker Attribution in Historical Italian Parliamentary Debates</i> Claudia Corbetta, Samuele Mazzei and Alessio Palmero Aprosio	65
<i>Computational Political Landscape of the Netherlands and Prime Minister Schoof's Position</i> Wessel Ledder and Iris Hendrickx	77

Workshop Program

16 May 2026

14:00–14:40 **Keynote**

From Concordance to Inference: ParlaCAP Helps ParlaMint Escape the Linguistics Lab

Nikola Ljubešić

14:40–15:40 **Linguistic and Discourse Analysis**

Quantifying Code-Switching in a Ukrainian Parliamentary Dataset 1990-2021

Olha Kanishcheva and Maria Shvedova

Ours and Yours: A Discourse Analysis of Political Identity Markers in Slovenian Parliamentary Discourse

Meden Katja

Representations of Europe and the European Union in Parliamentary Discourse from a Corpus-Assisted Perspective

Anna Kryvenko

15:40–16:00 **Corpus Creation I**

Towards ParlaMint-DE: Improving the Interoperability of the GermaParl Corpus of Plenary Protocols of the German Bundestag

Christoph Leonhardt and Andreas Blätte

16 May 2026 (continued)

16:40–17:00 Corpus Creation II

From Transcripts to Insights: A Digital Corpus and Interactive Speech Analysis Platform for Turkish Parliamentary Records

Basak Tepe, Irem Nur Yildirim, Onur Gungor and Susan Uskudarli

17:00–18:00 Processing of Parliamentary Data

Transcription and Recognition of Italian Parliamentary Speeches Using Vision-Language Models

Luigi Curini, Alfio Ferrara, Giovanni Pagano and Sergio Picascia

Beyond OCR: Structural Segmentation and Speaker Attribution in Historical Italian Parliamentary Debates

Claudia Corbetta, Samuele Mazzei and Alessio Palmero Aprosio

Computational Political Landscape of the Netherlands and Prime Minister Schoof's Position

Wessel Ledder and Iris Hendrickx