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

Nikola Ljubešić
Jožef Stefan Institute, University of Ljubljana,
Institute of Contemporary History, Ljubljana, Slovenia
nikola.ljubesic@ijs.si

Abstract

ParlaCAP is an OSCARS Open Science cascading grant project aimed at extending the use of the ParlaMint parliamentary corpora beyond corpus linguistics into the wider Social Sciences and Humanities (SSH). While ParlaMint provides a rich, comparable collection of parliamentary debates and accompanying metadata, its broader uptake has been limited. ParlaCAP addresses this by enriching the data with automatically derived political agendas and sentiment, enabling new forms of comparative political analysis. Using recent advances in multilingual transformer models, the project annotates over 8 million speeches from 28 European parliaments in more than 20 languages. By integrating ParlaMint with the Comparative Agendas Project (CAP) coding schema, ParlaCAP produces a FAIR dataset suitable for cross-national research on interaction of policy, sentiment, and political identity. The enrichments rely on two models, XLM-R-ParlaSent and XLM-R-ParlaCAP, both performing comparably to human annotators. The latter is trained using a teacher–student approach, where GPT-4o-generated labels are used to fine-tune a scalable classifier. The dataset is available via the CROSSDA repository and a user-friendly API. The talk concludes with a series of use cases demonstrating how meaningful insights can be obtained with minimal technical effort.

1. Summary of the Talk

I will present the results of ParlaCAP[1], an OSCARS Open Science cascading grant initiative aimed at extending the usability of the ParlaMint corpora beyond their traditional audience of corpus linguists. While ParlaMint[2] offers a rich and comparable collection of parliamentary debates across Europe, its uptake in broader Social Sciences and Humanities (SSH) fields, such as political science and sociology, has remained limited. ParlaCAP addresses this gap by transforming ParlaMint into a semantically enriched, analysis-ready dataset for comparative political research.

The project leverages recent advances in natural language processing and artificial intelligence to automatically identify political agendas and sentiments in debates from 28 European parliaments. The dataset comprises more than 8 million speeches in over 20 languages, making manual annotation infeasible. However, multilingual transformer models now enable highly consistent and accurate large-scale coding across languages and contexts.

A central contribution of ParlaCAP is the integration of ParlaMint with the Comparative Agendas Project (CAP) coding scheme. This allows for the automatic assignment of policy topics to parliamentary speeches, effectively bridging linguistic resources and political science methodologies. The result is a FAIR dataset that supports cross-national and longitudinal analyses of political agendas and enhances transparency in legislative discourse.

The enrichment is driven by two models: XLM-R-ParlaSent for sentiment analysis and XLM-R-ParlaCAP for agenda classification. Both are based on the XLM-RoBERTa architecture and perform comparably to human annotators. Notably, XLM-R-ParlaCAP is developed using a teacher–student approach, where the GPT-4o teacher generated labels that are then used to fine-tune the XLM-R student, combining the strengths of large language models with scalable deployment.

The resulting dataset is openly available via the Croatian CESSDA repository CROSSDA[3] and through a user-friendly API[4], that simplifies flexible data selection and followup analysis. Users can filter data by country, time, speaker attributes, agenda categories, and sentiment, supporting both exploratory and reproducible research.

The talk concludes with a series of use cases demonstrating how meaningful insights can be obtained with minimal technical effort, often in just a few lines of Python. These examples highlight shifts in policy attention, cross-country differences in political sentiment, and new opportunities for interdisciplinary research.

ParlaCAP thus helps ParlaMint "escape the linguistics lab", making parliamentary corpora accessible, interpretable, and valuable for a much broader SSH community.

[1] https://clarinsi.github.io/parlacap/
[2] https://www.clarin.eu/parlamint
[3] https://doi.org/10.23669/1ZTELP
[4] https://parlacap.ipipan.waw.pl

Proceedings of ParlaCLARIN V Workshop on Interoperability, Multilinguality, and Multimodality in Parliamentary Corpora @ LREC 2026, page 1
16 May 2026. ©ELRA Language Resources Association (ELRA), 2026