Put Them In to Get Them Out: the ParlaMint Corpora for Digital Humanities and Social Sciences Research

Fišer, Darja
Darja.Fiser@inz.si
Institute of Contemporary History, Slovenia

Kryvenko, Anna
Ganna.Kryvenko@inz.si
Institute of Contemporary History, Slovenia; NISS, Ukraine

Osenova, Petya
petya@bultreebank.org
Bulgarian Academy of Sciences, Bulgaria; Sofia University "St. Kl. Ohridski", Bulgaria

Pahor de Maiti, Kristina
Kristina.PahordeMaiti@ff.uni-lj.si
University of Ljubljana, Slovenia

1. Background and overview of the tutorial

Given a growing interest in the roles and discourses of national parliaments across Europe, records of parliamentary debates have recently become a fruitful object of study in a wide range of disciplines in social sciences and humanities, including but not limited to sociology (Skubic / Fišer 2022), history (Pančur / Šorn 2016; Jarlbrink / Norén 2022), political communication (Vliegenthart /Damstra 2019) and linguistic discourse analysis (Ilie 2016, Vouti-lainen 2017). However, the existing corpora of parliamentary de-bates are often incomparable in terms of timespan, metadata and linguistic annotation, which makes comparative research very dif-ficult if not impossible. To address this issue, the ParlaMint pro-ject was launched to make parliamentary corpora FAIR (findable, accessible, interoperable and reusable).

This half-day tutorial will introduce the audience to the ParlaMint corpora (Erjavec et al. 2023), which comprise records of 17 European national parliaments for the 2015–2019 period with almost half a billion words to date. Currently, the ParlaMint corpora are undergoing expansion with the 2021 and 2022 records for the existing parliaments and with records of additional national and regional parliaments not previously included. All these corpora are uniformly sampled, annotated and rich in comparable indivi-dual speaker and institutional group metadata including speaker name, gender, speaker role, affiliations with political parties and groups as well as coalitions or oppositions of political parties and so on. Changes in individual or group affiliations are also time-stamped if they occur in the time frame of the corpora. All the parliamentary records in the corpora are currently split into two datasets: Reference Subcorpus: from 2015 until end of 2019, and COVID Subcorpus: from 01.01.2020 onwards).

The proposed hands-on tutorial aims to motivate researchers in and outside Digital Humanities, who have no or little exposure to corpus analysis, to explore the potential of the ParlaMint corpora as an enriched and unified language resource for research into specific national parliaments and trans-nationally as well as to encourage cross-disciplinary cooperation.

2. Relevance to the Digital Humanities community

The ParlaMint corpora are openly available as downloadable data through the CLARIN repository either as plain corpora (http://hdl.handle.net/11356/1432) or corpora with added linguistic annotations (https://hdl.handle.net/11356/1431) as well as on GitHub (https://github.com/clarin-eric/ParlaMint), and through concordancers, e.g., (https://www.clarin.si/noske). The corpora can be directly queried through the concordancers or further processed through third-party text mining tools. The ParlaMint resource is designed as a sustainable initiative. Hence, the contributing partners are committed to schema and metadata improvements, corpus expansion and enrichment, as well as engagement activities (see Ogrodniczuk et al. 2022 for an overview).

3. Instructors

Petya Osenova is professor in morphology, syntax and corpus linguistics at the Sofia University "St. Kl. Ohridski" and senior researcher at the Institute of Information and Communication Technologies, Bulgarian Academy of Sciences. Her main interests are in the areas of language modeling, constraint-based linguistic theories, language resources creation, the lexicon-grammar interface, among others. Petya Osenova is part of the core team in the joint CLARIN and DARIAH infrastructure in Bulgaria – CLaDA-BG. She is the co-coordinator of the ParlaMint Project together with Maciej Ogrodniczuk from the Institute of Computer Science, Polish Academy of Sciences.

Anna Kryvenko has a PhD in linguistics and over 15 years of university teaching experience in the fields of lexical semantics, corpus analysis and discourse studies. Her research interests include corpus approaches to social studies and metaphor in political discourse. She is currently working on her postdoc MSCA Seal of Excellence project "The Changing Discursive Semantics of EU representations: Identity, Populism, Propaganda" funded by ARRS (Slovenia). She is a contributing partner for the ParlaMint-UA corpus.

Kristina Pahor de Maiti is a doctoral student in Linguistics at the University of Ljubljana and CY Cergy Paris University, and Research Assistant at the Institute of Contemporary History, Ljubljana. Her research interests include parliamentary discourse and computer-mediated communication, which she explores through the sociolinguistic and corpus-linguistic prism. In her doctoral research, she is focusing on the corpus-based analysis of socially unacceptable discourse online with a special emphasis on its figurative dimension.

4. Target audience and expected number of participants

This tutorial targets a broad range of scholars in Digital Humanities and social sciences interested in exploring the potential of parliamentary corpora as a language resource for studying socio-cultural phenomena. The showcases included in this tutorial might be of particular interest but are not limited to the European Studies community. No programming skills are needed and no prior experience in using language corpora or corpus querying tools is required to attend this tutorial.

Based on the experience of delivering a version of the *Voices of the Parliament: A Corpus Approach to Parliamentary Discourse Research* tutorial (https://sidih.si/20.500.12325/120), group size between 10 and 14 participants would be preferable.

5. Learning outcomes

Participants will learn how to formulate corpus queries, ranging from simple search strings for word forms and lemmata to complex CQL queries. Participants will be able to analyze the extracted structures in relation to the available speaker and text meta-

data. They will also be made aware of caution needed to ensure adequate framing and interpretation of the results for a wide range of research questions in Digital Humanities and Social Sciences including the speech and attitude of ruling parties in comparison to the opposition, specifics of protocol across parliaments, etc. The skills developed in this tutorial can be transferred to other types of corpora and concordancers. While the demo in Part 2 of the tutorial will be conducted on the ParlaMint-GB in order to ensure all participants can follow the queries and discuss the results, participants will be encouraged to select a ParlaMint corpus of their preference to carry out the hands-on activities in Part 3 of the tutorial.

6. Requirements for technical support

Conference organizers should provide an LCD projector, WIFI and a classroom suitable for pair work. Participants should bring their own laptops.

7. Timeline

(3.5 hours total incl. 30 min break)

- Part 1 – Introduction and demo session (1,5h):

20 min. - Introduction and overview of the ParlaMint project.

40 min. - ParlaMint corpus content including types of metadata and linguistic annotation, their comparability potential as well as parliament and language specific limitations to cross-corpus comparability.

30 min. - Demo in exploring the Crystal NoSketch Engine concordancer and using basic corpus analysis techniques to answer research questions related to the discursive construction of socially significant concepts in the ParlaMint-GB corpus over time.

- Break – 30 min.

- Part 2 – Hands-on session (1,5h):

30 min. - Participants exercise in creating subcorpora, performing simple and complex queries using regular expressions for concordance analysis as well as in keyword and collocation extraction in the ParlaMint-GB corpus.

10 min. - Participants create subcorpora in a ParlaMint corpus of their choice based on the types of the ParlaMint metadata, which correlate with their research interests.

15 min. - Participants extract keywords from the subcorpora created by them, compare and discuss their results in pairs.

35 min. - Participants extract collocates for one previously specified query from their subcorpora, discuss cross-corpus comparability of the results in pairs and report their findings to the rest of the group. Finally, participants give brief feedback on the applicability of the knowledge and skills acquired in this tutorial to their own research.

Bibliography

Erjavec, Tomaž / **Ogrodniczuk , Maciej** / **Osenova , Petya** / **Ljubešić Nikola** / **Simov, Kiril** / **Pan#ur, Andrej** / **Rudolf, Michał** / **Kopp, Matyáš** / **Barkarson, Starkaður** / **Steingrímsson, Steinþór** / **Çöltekin, Çagri** / **de Does, Jesse** / **Depuydt, Katrien** / **Agnoloni, Tommaso** / **Venturi, Giulia** / **Pérez, María Calzada** / **de Macedo, Luciana D.** / **Navarretta, Costanza** / **Luxardo, Giancarlo** / **Coole, Matthew** / **Rayson, Paul** / **Morkevičius, Vaidas** / **Krilavičius, Tomas** / **Darǵis, Roberts** / **Ring, Orsolya** / **van Heusden, Ruben** / **Marx, Maarten** / **Fišer, Darja** (2023): "The ParlaMint Corpora of Parliamentary Proceedings", in: *Language Resources and Evaluation* 57, 1: 415-448. DOI: 10.1007/s10579-021-09574-0 https://link.springer.com/article/10.1007/s10579-021-09574-0 [02.02.2023].

Ilie, Cornelia (2016): "Parliamentary Discourse and Deliberate Rhetoric", in: Ihalainen, Pasi / Ilie, Cornelia / Palonen, Kari (eds.): *Parliament and Parliamentarism: a Comparative History of a European Concept* . New York / Oxford: Berghahn 133-145.

Jarlbrink, Johan / **Norén, Frederik** (2022): "The rise and fall of 'propaganda' as a positive concept: a digital reading of Swedish parliamentary records, 1867-2019", in: *Scandinavian Journal of History* . DOI: 10.1080/03468755.2022.2134202 https://www.tandfonline.com/doi/full/10.1080/03468755.2022.2134202?scroll=top&needAccess=true&role=tab&aria-labelledby=full-article [03.11.2022].

Ogrodniczuk, Maciej / **Osenova, Petya** / **Erjavec, Tomaž** / **Fišer, Darja** / **Ljubešic, Nikola** / **Çöltekin, Çagri** / **Kopp, Matyáš** / **Meden, Katja** (2022): "ParlaMint II: The Show Must Go On", in: *Proceedings of the LREC 2022 ParlaCLARIN III Workshop on Creating, Enriching and Using Parliamentary Corpora* , Marseille, France, 20 June 2022: 1-6 http://www.lrec-conf.org/proceedings/lrec2022/workshops/ParlaCLARINIII/2022.parlaclariniii-1.0.pdf [03.11.2022].

Pančur, Andrej / **Šorn, Mojca** (2016): "Smart Big Data: Use of Slovenian Parliamentary Papers in Digital History", in: *Prispevki za novejšo zgodovino* 56, 3: 130-146 https://ojs.inz.si/pnz/article/view/193 [03.11.2022].

Skubic, Jure / **Fišer, Darja** (2022): "Parliamentary discourse research in sociology: Literature review", in: *Proceedings of the Third ParlaCLARIN Workshop*, Marseille, France, June 2022: 81-91 http://www.lrec-conf.org/proceedings/lrec2022/workshops/ParlaCLARINIII/pdf/2022.parlaclariniii-1.12.pdf [03.11.2022].

Vliegenthart, Rens / **Damstra, Alyt** (2019): "Parliamentary Questions, Newspaper Coverage, and Consumer Confidence in Times of Crisis: A Cross-National Comparison", in: *Political Communication* , 36, 1: 17-35. DOI: 10.1080/10584609.2018.1478472

Voutilainen, Eero (2017): „Parliamentary Records as Data for Linguistic Discourse Studies", in: *CLARIN-PLUS Workshop "Working with Parliamentary Records"* , Sofia, Bulgaria, 27-29 March 2017. https://www.clarin.eu/sites/default/files/1-voutilainen.pdf [03.11.2022].