ParlaMint: Comparable Corpora of European Parliamentary Data

Tomaž Erjavec
Jožef Stefan Institute, Slovenia
tomaz.erjavec@ijs.si

Maciej Ogrodniczuk
Institute of Computer Science PAS, Poland
maciej.ogrodniczuk@gmail.com

Petya Osenova
IICT-BAS, Bulgaria

Andrej Pančur
Institute of Contemporary History, Slovenia

Nikola Ljubešić
Jožef Stefan Institute, Slovenia

Tommaso Agnoloni
CNR-IGSG, Italy

Starkaður Barkarson
Árni Magnússon Institute for Icelandic Studies

María Calzada Pérez
Universitat Jaume I, Spain

Çağrı Çöltekin
University of Tübingen, Germany

Matthew Coole
Lancaster University, the UK

Roberts Darģis
IMCS UL, Latvia

Luciana D. de Macedo
Univ. Federal de Minas Gerais, Brazil

Jesse de Does
Dutch Language Institute,
the Netherlands

Katrien Depuydt
Dutch Language Institute,
the Netherlands

Sascha Diwersy
Univ. Paul Valéry Montpellier 3, France

Dorte Haltrup Hansen
University of Copenhagen, Denmark

Matyáš Kopp
Charles University, the Czech Republic

Tomas Krilavičius
Vytautas Magnus University, Lithuania

Giancarlo Luxardo
Univ. Paul Valéry Montpellier 3,
France

Maarten Marx
Universiteit van Amsterdam,
the Netherlands

Vaidas Morkevičius
Kaunas University of Technology, Lithuania

Costanza Navarretta
University of Copenhagen, Denmark

Paul Rayson
Lancaster University, the UK

Orsolya Ring
Centre for Social Sciences, Hungary

Michał Rudolf
Institute for Computer Science PAS, Poland

Kiril Simov
IICT-BAS, Bulgaria

Steinþór Steingrímsson
Árni Magnússon Institute for Icelandic Studies

István Üveges
University of Szeged, Hungary

Ruben van Heusden
Universiteit van Amsterdam, the Netherlands

Giulia Venturi
CNR-ILC, Italy

Abstract

This paper outlines the ParlaMint project from the perspective of its goals, tasks, participants, results and applications potential. The project produced language corpora from the sessions of the national parliaments of 17 countries, almost half a billion words in total. The corpora are split into COVID-related subcorpora (from November 2019) and reference corpora (to October 2019). The corpora are uniformly encoded according to the ParlaMint schema with the same Universal Dependencies linguistic annotations. Samples of the corpora and conversion scripts are available from the project's GitHub repository. The complete corpora are openly available via the CLARIN.SI repository[1] for download, and through the NoSketch Engine[2] and KonText[3] concordancers as well as through the Parlameter[4] interface for exploration and analysis.

1 Introduction

ParlaMint[5] (July 2020 – May 2021) was a project that built on the achievements of the ParlaCLARIN community and methodology and was financially supported by CLARIN-ERIC. The mission of ParlaMint was to turn existing contemporary diverse cross-national parliamentary data into resources that are comparable, interpretable and highly communicative with respect to society (NGOs, citizens, researchers, etc.). The ParlaMint project started with the creation of recent corpora of parliamentary sessions for 4 parliaments: Bulgarian, Croatian, Polish and Slovene. The project was then extended with 13 additional parliamentary corpora of the following countries: Belgium, the Czech Republic, Denmark, France, Hungary, Iceland, Italy, Latvia, Lithuania, the Netherlands, Turkey, and the UK. In addition, Spanish parliament data were added on a voluntary basis.

The project aimed to provide data and tools for focused observations on trends, opinions, decisions on lock-downs and restrictive measures as well as on the consequences with respect to health, medical care systems, employment, etc. in times of emergencies. For the ParlaMint project the emergency case is obvious – the COVID-19 pandemic. However, the methodology is scalable also to other events, such as economic crises, etc. Thus, the main aims of the project were: to compile a collection of parliamentary corpora from a number of countries and in a number of languages in a harmonized format, covering both current data and older, reference data; to process the corpora linguistically; to index the data with popular concordancers so that interested parties can search and extract the relevant comparable information; to make the data, workflow descriptions, related standards and lessons learnt publicly available; to show through appropriate use cases that the CLARIN resources and technology serve societal needs.

Considerable effort was already put into data from European Parliament, so we have at disposal valuable and well-synchronized resources like Europarl (Koehn, 2005),[6] JRC-Acquis (Steinberger et al., 2006)[7] or DCEP: Digital Corpus of the European Parliament (Hajlaoui et al., 2014).[8]

At the same time, there are many ongoing national initiatives ranging from parliament-focused corpora to task-oriented ones. Within large EU initiatives, such as CLARIN-ERIC, identification was performed of the available resources within European countries. It is worth mentioning that parliamentary data were one of the CLARIN Key Resource Families (Fišer et al., 2018).[9]

A number of related workshops have also been organized on the topics of gathering, standardizing, processing, maintaining, visualizing and using parliamentary data, in particular: CLARIN-PLUS Work-

[1] https://www.clarin.si/repository/xmlui/handle/11356/1432 and https://www.clarin.si/repository/xmlui/handle/11356/1431
[2] http://www.clarin.si/noske/
[3] https://www.clarin.si/kontext/corpora/corplist
[4] https://parlamint.parlameter.org/poslanske-skupine
[5] https://www.clarin.eu/content/parlamint
[6] http://www.statmt.org/europarl/
[7] https://ec.europa.eu/jrc/en/language-technologies/jrc-acquis
[8] https://ec.europa.eu/jrc/en/language-technologies/dcep
[9] https://www.clarin.eu/resource-families/parliamentary-corpora

shop "Working with Parliamentary Records"[10] (2017); two ParlaCLARIN workshops at LREC 2018 and 2020 (Fišer et al., 2018; Fišer et al., 2020)[11] or CLARIN Interoperability Committee ParlaFormat workshop[12] (2019) on standardization of parliamentary data.

Parliamentary data have also been subject of growing interest of the digital humanities reflected in search for synergies with the natural language processing community. This resulted in such events as *Computational Analysis of Political Texts* tutorial[13] offered at the top venues of computational social science and natural language processing (IC2S2 2019[14] and ACL 2019[15]) or *Big Data and the Study of Language and Culture: Parliamentary Discourse across Time and Space* workshop[16].

The paper is organized as follows: in the next section the structure and availability of the ParlaMint corpora is outlined. Section 3 briefly showcases the participating languages and parliaments. Section 4 concludes the paper.

2 Structure and availability of the corpora

ParlaMint contains 17 corpora with 16 languages (the Belgian corpus is bilingual Dutch/French), and comprises 22 thousand files, over 3.5 million speeches and almost 500 million words. It defines over 11 thousand persons and over 1.5 thousand "organisations", i.e. political parties, parliamentary groups etc.

In Figure 1 we give an overview of the ParlaMint corpora. The left side gives the time period covered by each corpus, with the dashed line (November 2019) splitting the period into "reference" and COVID-19 subcorpora. The middle part gives the country code, and the right part shows the number of words contained in the corpora. As can be seen, most corpora start in 2015, with the earliest speeches from 2009, and, while most corpora end mid-2020, the latest extends to April 2021. As for sizes, by far the largest corpus, both per year and in total, is that of the UK, with even the fact that it contains the speeches of both the House of Lords and of the House of Commons not fully explaining its size, but must be (as it is with the French) a result of longer or more sessions of their parliaments. In the opposite direction, the outlier is the Hungarian corpus, where its small size is due to the fact that it contains only interpellations and urgent questions from plenary sessions of the parliament.

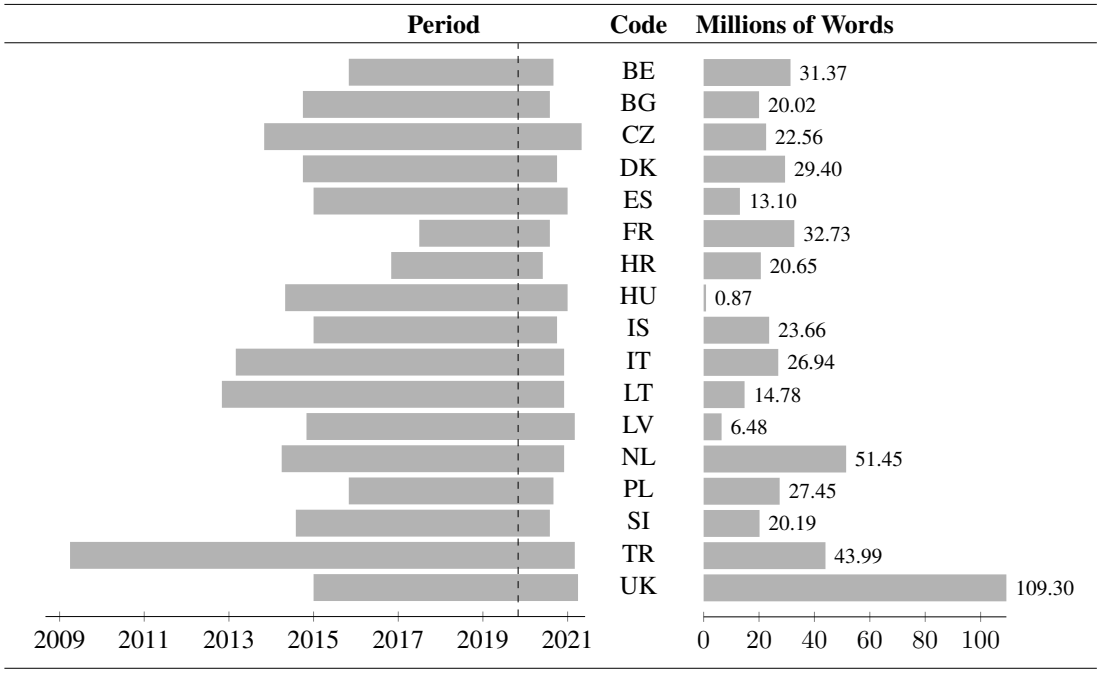
The corpora have extensive metadata about the speakers (speaker name, gender, party affiliation, MP status). They are structured into time-stamped terms, sessions and meetings, with each speech being marked by its speaker and their role (chair, regular speaker). The speeches contain also marked-up transcriber comments, such as gaps in the transcription, interruptions, applause, etc.

The corpora are encoded according to the Parla-CLARIN TEI recommendation[17] but have been validated to conform to the much stricter ParlaMint schemas, available from the ParlaMint GitHub repository.[18] This repository includes, apart from the XML schemas, also content validation scripts, scripts to convert the corpora into other formats, as well as samples from all the available corpora.

The corpora are available under CC BY via the CLARIN.SI repository in two variants, the "plain text" (Erjavec et al., 2021b) and the linguistically annotated one (Erjavec et al., 2021a). The former includes all the metadata and structured transcription in XML and in derived plain text format, while the latter adds linguistic annotations, which include named entities, lemmatisation, morphological features and syntactic parses according to the Universal Dependencies recommendations.[19] This version also includes the corpora in derived CoNLL-U and so called vertical formats. Samples of the "plain text" and linguistically annotated corpora, as well as the samples in several derived formats are also available from the ParlaMint GitHub repository.

[10] https://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records
[11] https://www.clarin.eu/ParlaCLARIN, https://www.clarin.eu/ParlaCLARIN-II
[12] https://www.clarin.eu/event/2019/parlaformat-workshop
[13] https://poltexttutorial.wordpress.com/
[14] 5th International Conference on Computational Social Science, https://2019.ic2s2.org/
[15] 57th Annual Meeting of the Association for Computational Linguistics, https://acl2019.org/
[16] Collocated with 40th Intl. Computer Archive of Modern and Medieval English conference, http://icame.uib.no/
[17] https://clarin-eric.github.io/parla-clarin
[18] https://github.com/clarin-eric/ParlaMint
[19] https://universaldependencies.org

Figure 1: The time period and number of words of the ParlaMint corpora.



3 Compilation of the ParlaMint corpora

The corpora of the following countries are included in ParlaMint: Belgium, Bulgaria, Croatia, the Czech Republic, Denmark, France, Hungary, Iceland, Italy, Latvia, Lithuania, Poland, Spain, the Netherlands, Slovenia, Turkey and the UK.

First of all, these countries have different political and thus, parliamentary systems. For example, there are unicameral (Bulgaria, Croatia, Denmark, Hungary, Iceland, Latvia, Lithuania, Turkey) and bicameral parliaments (Belgium, the Czech Republic, France, Italy, Poland, Slovenia, Spain, the Netherlands, the UK), each with its own specifics, which is reflected in the structure of the particular corpora, e.g. whether they distinguish sessions, sittings, and meetings. The steps of getting the data, converting them to the ParlaMint schema and annotating it linguistically also varied across the corpora.

Getting the data required either scraping it from the parliamentary websites (Belgium, Bulgaria, the Czech Republic, Hungary, Iceland, Latvia, Spain, Turkey); obtaining via Parlameter API (Croatia); retrieving from an already maintained parliamentary corpus (Poland and Slovenia); downloading from a server (Denmark, France, the Netherlands); obtaining through parliamentary API (UK) or through a service center at the parliament (Italy).

Data conversion employed various strategies such as: incremental and semi-automatic transformation from HTML to basic TEI XML and then to the ParlaMint format through XML constraints (Bulgarian) or through XSLT stylesheets and Python, Perl and Bash scripts (Belgian, Dutch, French, Spanish); automatic conversion through Perl scripts with heuristics only for difficult parts such as the transcriber comments (Croatian, Czech, Danish); automatic conversion through Python scripts with possible corrections of data during the process (Hungarian, Icelandic, Latvian, Polish, Turkish); transformation with XSLT, and some manual interventions upstream (Slovene) or adding necessary extensions to XSLT (English); automatic conversion with JAVA code (Italian). The main challenges of the conversion were related to re-structuring the data, and esp. adding mark-up to the previously unstructured data.

Linguistic processing included the UD-based morphosyntactic annotation and a named entity annotation with the traditional NEs: Person, Location, Organization and Misc.

This step was also approached differently by the groups depending on the availability of these tools for the language and their quality and performance. Thus, for some languages pre-trained pipelines were used that follow the same model. For example, the CLASSLA pipeline[20] was used for the annotation of Bulgarian, Croatian, and Slovene corpora. Italian, French and Spanish relied on the Stanza NLP pipeline, while for English the Stanford NLP pipeline was used. In the Spanish case, the Stanza NLP pipeline was aided by AnCora Treebanks and corpora.[21]

Other languages used language-specific models either different for each step, or in a combined piped mode, which was the case for Belgian, Czech, Danish, Dutch, Hungarian, Icelandic, Latvian, and Polish.

Some corpora contain additional linguistic information, e.g. Croatian and Slovene have also the MULTEXT-East (Erjavec, 2012) morphosyntactic annotations, while Czech also contains their own highly detailed and nested NE annotations.

4 Conclusions

The ParlaMint project establishes an innovative strategy for handling parliamentary data and processing them in times of any emergency period (COVID-19 is just a showcase). The novelties relate to unified handling of cross-lingual and cross-parliament comparable data, and to the quick access of all interested parties to these data.

The project output was already used in several studies. ParlaMint took part in the Helsinki Digital Humanities Hackathon DHH21 (19–28.05.2021).[22] The corpora were explored in three practical showcases: on Science and Expertise in Parliaments[23] on a comparative analysis of the available corpora[24] and on the Parlameter service of the ParlaMint project.[25]

The Parla-CLARIN TEI encoding is becoming a de-facto standard for national parliamentary data, and it will be further developed to cover more detailed and specific metadata across languages and parliaments. The created openly available corpora can serve as a baseline for further updates. Such uniform updates across the corpora would strongly support various methods of comparative research across parliaments and political systems.

We believe that the availability of comparable multilingual parliamentary data will boost further the research in the areas of digital humanities, linguistics, politology, sociology, psychology as well as in all the related branches of sciences.

Acknowledgements

We would like to thank CLARIN-ERIC for the financial support of ParlaMint.

The work on Bulgarian Parliamentary data was partially supported by the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH – CLaDA-BG, Grant number DO1-377/18.12.2020.

The work on the Czech Parliamentary data was partially supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2018101 LINDAT/CLARIAH-CZ.

The work on the Danish ParlaMint corpus was partially supported by the Department of Nordic Studies and Linguistics at the University of Copenhagen through CLARIN-DK.

The work on Hungarian Parliamentary data was partially supported by the Ministry of Innovation and Technology NRDI Office within the framework of the Artificial Intelligence National Laboratory Program, No. NKFIH-870-8/2020; and received funding from the European Union's Horizon 2020 research and innovation programme under Grant Agreement No. 951832 (OPTED).

[20] `https://pypi.org/project/classla/`
[21] `https://universaldependencies.org/treebanks/es_ancora/index.html`
[22] `https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/`
[23] See the video: `https://www.youtube.com/watch?v=K4y03qr4WoU`
[24] See the video: `https://www.youtube.com/watch?v=ddBHvbuzke4`
[25] See the video: `https://www.youtube.com/watch?v=h1292E_vtO8`

The work on Polish parliamentary data was partially supported by project CESAR (CEntral and South-east europeAn Resources, a European CIP ICT-PSP project, grant agreement 271022), CLARIN-PL (a Polish Ministry of Science and Education project, grant numbers DIR/WK/2016/02 and DIR/WK/2018/01) and MARCELL (Multilingual Resources for CEF.AT in the legal domain, a CEF-TC-2017-3 – eTranslation grant, grant agreement INEA/CEF/ICT/A2017/1565710, co-financed by the Polish Ministry of Science and Higher Education: research project 4082/CEF/2018/2, funds for 2018–2020).

The work on the Spanish Parliamentary corpus was supported by the Spanish Ministry of Science and Innovation, PID2019-108866RB-I0 / AEI /10.13039/501100011033, "Original, translated and interpreted representations of the refugee cris(e)s: methodological triangulation within corpus-based discourse studies".

The work on the Latvian Parliamentary data was partially supported by the CLARIN-LV, European Regional Development Fund project "University of Latvia and institutes in the European Research Area – Excellency, activity, mobility, capacity" (1.1.1.5/18/I/016) and Latvian State Research Programme's project "Digital Resources for Humanities: Integration and Developmen" (VPP-IZM-DH-2020/1-0001).

The work on the Slovenian Parliamentary data was partially supported by the Research infrastructures CLARIN.SI and DARIAH-SI, and the Slovenian Research Agency research programme P2-103 "Knowledge Technologies".

We thank Mindaugas Petkevičius, Monika Briedienė and Andrius Utka for their help in creating the Lithuanian corpora.

We thank Bart Jongejan who contributed to the creation of the Danish corpus.

References

Tomaž Erjavec et al. 2021a. *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1431`.

Tomaž Erjavec et al. 2021b. *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*. Slovenian language resource repository CLARIN.SI. `http://hdl.handle.net/11356/1432`.

Tomaž Erjavec. 2012. MULTEXT-East: Morphosyntactic Resources for Central and Eastern European Languages. *Language Resources and Evaluation*, 46(1):131–142.

Darja Fišer, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Darja Fišer, Maria Eskevich, and Franciska de Jong, editors. 2020. *Proceedings of the Second ParlaCLARIN Workshop*, Marseille, France. European Language Resources Association (ELRA).

Darja Fišer, Maria Eskevich, and Franciska de Jong, editors. 2018. *Proceedings of LREC 2018 Workshop ParlaCLARIN: Creating and Using Parliamentary Corpora*, Paris, France. European Language Resources Association (ELRA).

Najeh Hajlaoui, David Kolovratnik, Jaakko Väyrynen, Ralf Steinberger, and Daniel Varga. 2014. DCEP – Digital Corpus of the European Parliament. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the Tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Ralf Steinberger, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiș, and Dániel Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *CoRR*, abs/cs/0609058.