

CLARIN Annual Conference 2021

PROCEEDINGS

Edited by

Monica Monachini, Maria Eskevich

27 – 29 September 2021
Virtual Edition

Please cite as:
Proceedings of CLARIN Annual Conference 2021. Eds. M. Monachini and M. Eskevich.
Virtual Edition, 2021.

Programme Committee

Chair:

- Monica Monachini, Institute of Computational Linguistics "A. Zampolli" (IT)

Members:

- Lars Borin, University of Gothenburg (SE)
- António Branco, Universidade de Lisboa (PT)
- Tomaž Erjavec, Jožef Stefan Institute (SI)
- Eva Hajičová, Charles University Prague (CZ)
- Erhard Hinrichs, University of Tübingen (DE)
- Marinos Ioannides, Cyprus University of Technology (CY)
- Langa Khumalo, North West University (ZA)
- Nicolas Larrousse, Huma-Num (FR)
- Krister Lindén, University of Helsinki (FI)
- Karlheinz Mörth, Austrian Academy of Sciences (AT)
- Costanza Navarretta, University of Copenhagen (DK)
- Jan Odijk, Utrecht University (NL)
- Maciej Piasecki, Wrocław University of Science and Technology (PL)
- Stelios Piperidis, Institute for Language and Speech Processing (ILSP), Athena Research Center (GR)
- Eiríkur Rögnvaldsson, University of Iceland (IS)
- Kiril Simov, IICT, Bulgarian Academy of Sciences (BG)
- Inguna Skadiņa, University of Latvia (LV)
- Koenraad De Smedt, University of Bergen (NO)
- Marko Tadić, University of Zagreb (HR)
- Jurgita Vaičenonienė, Vytautas Magnus University (LT)
- Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences (HU)
- Kadri Vider, University of Tartu (EE)
- Martin Wynne, University of Oxford (UK)

Reviewers:

- Lars Borin, SE
- António Branco, PT
- Tomaž Erjavec, SI
- Eva Hajičová, CZ
- Martin Hennelly, ZA
- Erhard Hinrichs, DE
- Marinos Ioannides, CY
- Nicolas Larrousse, FR
- Krister Lindén, FI
- Monica Monachini, IT
- Karlheinz Mörth, AT
- Costanza Navarretta, DK
- Jan Odijk, NL
- Stelios Piperidis, GR
- Eiríkur Rögnvaldsson, IS
- Kiril Simov, BG
- Inguna Skadiņa, LV
- Koenraad De Smedt, NO
- Marko Tadić, HR
- Jurgita Vaičėnienė, LT
- Tamás Váradi, HU
- Kadri Vider, EE
- Martin Wynne, UK

Subreviewers:

- Federico Boschetti, IT
- Christophe Parisse, FR
- Thorsten Trippel, DE
- Valeria Quochi, IT
- Zijian Győző Yang, HU
- Efstathia Soroli, FR
- Enikő Héja, HU
- Bence Nyéki, HU
- Angelo Mario Del Grosso, IT
- Olivier Baude, FR
- Kinga Jelencsik-Mátyus, HU

CLARIN 2021 submissions, review process and acceptance

- Call for abstracts: 19 January 2021, 1 March 2021
- Submission deadline: 28 April 2021
- In total 40 submissions were received and reviewed (three reviews per submission)
- Virtual PC meeting: 16-17 June 2021
- Notifications to authors: 22 June 2021
- 35 accepted submissions

More details on the paper selection procedure and the conference can be found at <https://www.clarin.eu/event/2021/clarin-annual-conference-2021-virtual-event>.

Table of Contents

Research cases

<i>How to Perform Linguistic Analysis of Emotions in a Corpus of Vernacular Semiliterate Speech with the Help of CLARIN Tools</i> Rosalba Nodari and Luisa Corona	1
<i>Dependency Trees in Automatic Inflection of Multi Word Expressions in Polish</i> Ryszard Tuora and Łukasz Kobyliński	6
<i>Corpora for Bilingual Terminology Extraction in Cybersecurity Domain</i> Andrius Utkā, Sigita Rackevičienė, Liudmila Mockienė, Aivaras Rokas, Marius Laurinaitis and Agnė Bielinskienė	11

Resources

<i>Voices from Ravensbrück. Towards the Creation of an Oral and Multi-lingual Resource Family</i> Silvia Calamai, Jeannine Beeken, Henk Van Den Heuvel, Max Broekhuizen, Arjan van Hessen, Christoph Draxler and Stefania Scagliola	16
<i>ParlaMint: Comparable Corpora of European Parliamentary Data</i> Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova	20
<i>The Nature of Icelandic as a Second Language: An Insight from the Learner Error Corpus for Icelandic</i> Isidora Glisic and Anton Karl Ingason	26
<i>Insights on a Swedish Covid-19 Corpus</i> Dimitrios Kokkinakis	48
<i>From Data Collection to Data Archiving: A Corpus of Italian Spontaneous Speech</i> Daniela Mereu	35
<i>IceTaboo: A Database of Contextually Inappropriate Words for Icelandic</i> Agnė Sólmundsdóttir, Lilja Björk Stefánsdóttir and Anton Karl Ingason	39
<i>The CIRCSE Collection of Linguistic Resources in CLARIN-IT</i> Rachele Sprugnoli and Marco Passarotti	44

<i>'Cretan Institutional Inscriptions' Meets CLARIN-IT</i> Irene Vagionakis, Riccardo Del Gratta, Federico Boschetti, Paola Baroni, Angelo Mario Del Grosso, Tiziana Mancinelli and Monica Monachini	48
--	----

<i>Swedish Word Metrics: A Swe-Clarín resource for Psycholinguistic Research in the Swedish Language</i> Erik Witte, Jens Edlund, Arne Jönsson and Henrik Danielsson	54
---	----

Annotation and Acquisition Tools

<i>Creating an Error Corpus: Annotation and Applicability</i> Þórunn Arnardóttir, Xindan Xu, Dagbjört Guðmundsdóttir, Lilja Björk Stefánsdóttir and Anton Karl Ingason	59
--	----

<i>ALEXIA: A Lexicon Acquisition Tool</i> Steinunn Rut Friðriksdóttir, Atli Jasonarson, Steinþór Steingrímsson and Einar Freyr Sigurðsson	64
--	----

<i>CLARIN Knowledge Centre for Belarusian Text and Speech Processing (K-BLP)</i> Yuras Hetsevich, Jauheniya Zianouka, David Latyshevich, Mikita Suprunchuk, Valer Varanovich and Katerina Lomat	68
---	----

<i>Enhancing CLARIN-DK Resources While Building the Danish ParlaMint Corpus</i> Bart Jongejan, Dorte Haltrup Hansen and Costanza Navarretta	73
--	----

<i>Annotation Management Tool: A Requirement for Corpus Construction</i> Yousuf Ali Mohammed, Arild Matsson and Elena Volodina	77
---	----

<i>A Method for Building Non-English Corpora for Abstractive Text Summarization</i> Julius Monsen and Arne Jönsson	82
---	----

<i>Reliability of Automatic Linguistic Annotation: Native vs Non-native Texts</i> Elena Volodina, David Alfter, Therese Lindström Tiedemann, Maisa Lauriala and Daniela Piipponen	90
--	----

Research Data Management, Metadata and Curation

<i>Seamless Integration of Continuous Quality Control and Research Data Management for Indigenous Language Resources</i> Anne Ferger and Daniel Jettka	95
---	----

<i>The TEI-based ISO Standard "Transcription of Spoken Language" as an Exchange Format within CLARIN and beyond</i> Hanna Hedeland and Thomas Schmidt	100
--	-----

<i>Curation Criteria for Multimodal and Multilingual Data: A Mixed Study within the Quest Project</i> Amy Isard and Elena Arestau	105
--	-----

<i>Flexible Metadata Schemes for Research Data repositories - The Common Framework in Dataverse and the CMDI Use Case</i>	
---	--

Jerry de Vries, Vyacheslav Tykhonov, Andrea Scharnhorst, Eko Indarto and Femmy Admiraal .	109
<i>Citation Tracking and Versioning for Linguistic Examples</i> Tobias Weber	114
<i>Bagman – A Tool that Supports Researchers Archiving Their Data</i> Claus Zinn	119
Repositories and National CLARIN Centres	
<i>Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS</i> Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Jón Guðnason, Þorsteinn Daði Gunnarsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Einar Freyr Sigurðsson, Atli Þór Sigurgeirsson, Vésteinn Snæbjarnarson and Steinþór Steingrímsson	124
<i>CLARIN-IT Resources in CLARIN ERIC - a Bird's-Eye View</i> Dario Del Fante, Francesca Frontini, Monica Monachini and Valeria Quochi	129
<i>A Data Repository for the Management of Dynamic Linguistic Datasets</i> Thomas Gaillat, Leonardo Contreras Roa and Juvéanal Attoumbre	134
<i>Opening Language Resource Infrastructures to Non-research Partners: Practicalities and Challenges</i> Verena Lyding, Egon W. Stemle and Alexander König	139
<i>CLARIN Flanders: New Prospects</i> Vincent Vandeghinste, Els Lefever, Walter Daelemans, Tim Van de Cruys and Sally Chambers ..	86
<i>ARCHE Suite: A Flexible Approach to Repository Metadata Management</i> Mateusz Żóltak, Martina Trognitz and Matej Durco	145
Legal Issues Related to the Use of LRs in Research	
<i>Legal Issues Related to the Use of Twitter Data in Language Research</i> Paweł Kamocki, Vanessa Hanneschläger, Esther Hoorn, Aleksei Kelli, Marc Kupietz, Krister Linden and Andrius Puksas	150
<i>The Interplay of Legal Regimes of Personal Data, Intellectual Property and Freedom of Expression in Language Research</i> Aleksei Kelli, Krister Lindén, Paweł Kamocki, Kadri Vider, Penny Labropoulou, Ramūnas Birvtonas, Vadim Mantrov, Vanessa Hanneschläger, Riccardo del Gratta, Age Värvi, Gaabriel Tavits and Andres Vutt	154
<i>Ethnomusicological Archives and Copyright Issues: an Italian Case Study</i> Prospero Marra, Duccio Piccardi and Silvia Calamai	160
<i>Less Is More when FAIR. The Minimum Level of Description in Pathological Oral and Written Data</i> Rosalba Nodari, Silvia Calamai and Henk van den Heuvel	166

ParlaMint: Comparable Corpora of European Parliamentary Data

<p>Tomaž Erjavec Jožef Stefan Institute, Slovenia tomaz.erjavec@ijs.si</p>	<p>Maciej Ogrodniczuk Institute of Computer Science PAS, Poland maciej.ogrodniczuk@gmail.com</p>
<p>Petya Osenova IICT-BAS, Bulgaria</p>	<p>Andrej Pančur Institute of Contemporary History, Slovenia</p>
<p>Nikola Ljubešić Jožef Stefan Institute, Slovenia</p>	<p>Tommaso Agnoloni CNR-IGSG, Italy</p>
<p>Starkaður Barkarson Árni Magnússon Institute for Icelandic Studies</p>	<p>María Calzada Pérez Universitat Jaume I, Spain</p>
<p>Çağrı Çöltekin University of Tübingen, Germany</p>	<p>Matthew Coole Lancaster University, the UK</p>
<p>Roberts Dargis IMCS UL, Latvia</p>	<p>Luciana D. de Macedo Univ. Federal de Minas Gerais, Brazil</p>
<p>Jesse de Does Dutch Language Institute, the Netherlands</p>	<p>Katrien Depuydt Dutch Language Institute, the Netherlands</p>
<p>Sascha Diwersy Univ. Paul Valéry Montpellier 3, France</p>	<p>Dorte Haltrup Hansen University of Copenhagen, Denmark</p>
<p>Matyáš Kopp Charles University, the Czech Republic</p>	<p>Tomas Krilavičius Vytautas Magnus University, Lithuania</p>
<p>Giancarlo Luxardo Univ. Paul Valéry Montpellier 3, France</p>	<p>Maarten Marx Universiteit van Amsterdam, the Netherlands</p>
<p>Vaidas Morkevičius Kaunas University of Technology, Lithuania</p>	<p>Costanza Navarretta University of Copenhagen, Denmark</p>
<p>Paul Rayson Lancaster University, the UK</p>	<p>Orsolya Ring Centre for Social Sciences, Hungary</p>
<p>Michał Rudolf Institute for Computer Science PAS, Poland</p>	<p>Kiril Simov IICT-BAS, Bulgaria</p>
<p>Steinþór Steingrímsson Árni Magnússon Institute for Icelandic Studies</p>	<p>István Üveges University of Szeged, Hungary</p>
<p>Ruben van Heusden Universiteit van Amsterdam, the Netherlands</p>	<p>Giulia Venturi CNR-ILC, Italy</p>

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 repository¹ for download, and through the NoSketch Engine² and KonText³ concordancers as well as through the ParlaMeter⁴ interface for exploration and analysis.

1 Introduction

ParlaMint⁵ (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-

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Licence details: <http://creativecommons.org/licenses/by/4.0/>

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.parlamer.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 workshop¹² (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* tutorial¹³ offered at the top venues of computational social science and natural language processing (IC2S2 2019¹⁴ and ACL 2019¹⁵) or *Big Data and the Study of Language and Culture: Parliamentary Discourse across Time and Space* workshop¹⁶.

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 recommendation¹⁷ 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/>

¹⁴5th International Conference on Computational Social Science, <https://2019.ic2s2.org/>

¹⁵57th Annual Meeting of the Association for Computational Linguistics, <https://acl2019.org/>

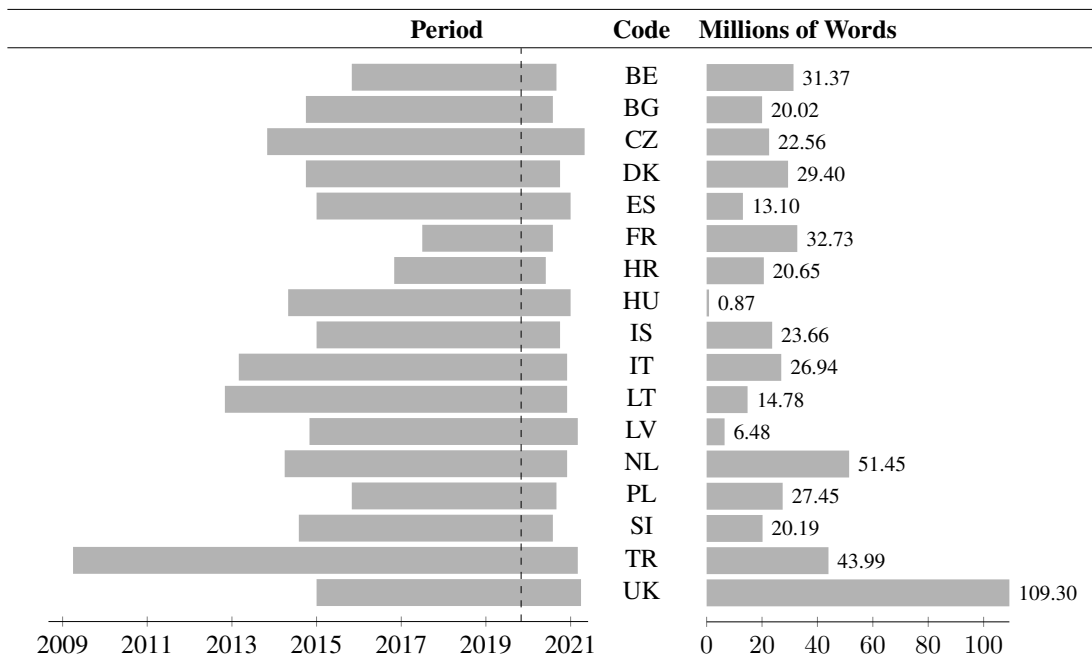
¹⁶Collocated with 40th Intl. Computer Archive of Modern and Medieval English conference, <http://icame.uib.no/>

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 pipeline²⁰ 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 show-cases: on Science and Expertise in Parliaments²³ on a comparative analysis of the available corpora²⁴ 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 DOI-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 NRD 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/>

²¹https://universaldependencies.org/treebanks/es_ancora/index.html

²²<https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/>

²³See the video: <https://www.youtube.com/watch?v=K4y03qr4WoU>

²⁴See the video: <https://www.youtube.com/watch?v=ddBHvbuzke4>

²⁵See the video: https://www.youtube.com/watch?v=h1292E_vt08

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 crisis: 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 Development” (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 Tufiş, and Dániel Varga. 2006. The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages. *CoRR*, abs/cs/0609058.