ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe

Tomaž Erjavec

tomaz.erjavec@ijs.si

Jožef Stefan Institute

Matyáš Kopp

Charles University

Nikola Ljubešić

Jožef Stefan Institute

Taja Kuzman

Jožef Stefan Institute

Paul Rayson

Lancaster University

Petya Osenova

Institute of Information and Communication Technologies

Maciej Ogrodniczuk

Institute of Computer Science

Çağrı Çöltekin

University of Tübingen

Danijel Koržinek

Polish-Japanese Academy of Information Technology

Katja Meden

Jožef Stefan Institute

Jure Skubic

Institute of Contemporary History

Peter Rupnik

Jožef Stefan Institute

Tommaso Agnoloni

Institute of Legal Informatics and Judicial Systems

José Aires

University of Lisbon

Starkaður Barkarson

Árni Magnússon Institute for Icelandic Studies

Roberto Bartolini
Institute for Computational Linguistics “A. Zampolli”

Núria Bel
Pompeu Fabra University

María Calzada Pérez
Jaume I University

Roberts Darģis
IMCS at the University of Latvia

Sascha Diwersy
Université Paul-Valéry Montpellier

Maria Gavriilidou
Athena Research Center

Ruben van Heusden
University of Amsterdam

Mikel Iruskieta
University of the Basque Country

Neeme Kahusk
University of Tartu

Anna Kryvenko
Institute of Contemporary History

Noémi Ligeti-Nagy
Hungarian Research Centre for Linguistics

Carmen Magariños
University of Santiago de Compostela

Martin Mölder
University of Tartu

Costanza Navarretta
University of Copenhagen

Kiril Simov
Institute of Information and Communication Technologies

Lars Magne Tungland
National Library of Norway

Jouni Tuominen
University of Helsinki

John Vidler
Lancaster University

Adina Ioana Vladu
University of Santiago de Compostela

Tanja Wissik

Austrian Centre for Digital Humanities and Cultural Heritage

Väinö Yrjänäinen

Uppsala University

Darja Fišer

Institute of Contemporary History

Research Article

Keywords: Parliamentary proceedings, Comparable corpora, TEI

Posted Date: April 1st, 2024

DOI: https://doi.org/10.21203/rs.3.rs-4176128/v1

Additional Declarations: No competing interests reported.

ParlaMint II: Advancing Comparable Parliamentary Corpora Across Europe

Tomaž Erjavec[1*], Matyáš Kopp[2*], Nikola Ljubešić[1,9], Taja Kuzman[1], Paul Rayson[3], Petya Osenova[4*], Maciej Ogrodniczuk[5], Çağrı Çöltekin[6], Danijel Koržinek[7], Katja Meden[1,8,9], Jure Skubic[9], Peter Rupnik[1], Tommaso Agnoloni[10], José Aires[11], Starkaður Barkarson[12], Roberto Bartolini[13], Núria Bel[14], María Calzada Pérez[15], Roberts Darģis[16], Sascha Diwersy[17], Maria Gavriilidou[18], Ruben van Heusden[19], Mikel Iruskieta[20], Neeme Kahusk[21], Anna Kryvenko[9,22], Noémi Ligeti-Nagy[23], Carmen Magariños[24], Martin Mölder[25], Costanza Navarretta[26], Kiril Simov[4], Lars Magne Tungland[27], Jouni Tuominen[28], John Vidler[3], Adina Ioana Vladu[24], Tanja Wissik[29], Väinö Yrjänäinen[30], and Darja Fišer[9]

[1]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana, Slovenia.

[2]Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University, Prague, Czech Republic.

[3]UCREL NLP research group, Lancaster University, Lancaster, UK.

[4]Institute of Information and Communication Technologies, Bulgarian Academy of Sciences, Sofia, Bulgaria.

[5]Institute of Computer Science, Polish Academy of Sciences, Warsaw, Poland.

[6]University of Tübingen, Tübingen, Germany.

[7]Department of Multimedia, Polish-Japanese Academy of Information Technology, Warsaw, Poland.

[8]Jožef Stefan International Postgraduate School, Ljubljana, Slovenia.

[9]Institute of Contemporary History, Ljubljana, Slovenia.

[10]Institute of Legal Informatics and Judicial Systems, CNR, Firenze, Italy.

[11]School of Arts and Humanities - Centre of Linguistics, University of Lisbon, Lisbon, Portugal.
[12]Department of Icelandic, The Árni Magnússon Institute for Icelandic Studies, Reykjavík, Iceland.
[13]Istituto di Linguistica Computazionale, CNR, Pisa, Italy.
[14]Department of Translation and Language Sciences, Pompeu Fabra University, Barcelona, Spain.
[15]Departamento de Traducción y Comunicación , Universitat Jaume I, Castellón, Spain.
[16]IMCS at the University of Latvia, Riga, Latvia.
[17]Praxiling UMR 5267 CNRS, Paul Valéry University Montpellier 3, Montpellier, France.
[18] Athena Research & Innovation Center in Information Communication & Knowledge Technologies, Athens, Greece.
[19]Information Retrieval Lab, University of Amsterdam, Amsterdam, The Netherlands.
[20]HiTZ Basque Center for Language Techonology, Ixa, University of the Basque Country (UPV/EHU), Bilbao, Spain.
[21]Institute of Computer Science, University of Tartu, Tartu, Estonia.
[22]NISS, Kyiv, Ukraine.
[23]Language Technology Research Group, HUN-REN Hungarian Research Centre for Linguistics, Budapest, Hungary.
[24]Galician Language Institute, University of Santiago de Compostela, Santiago de Compostela, Spain.
[25]Johan Skytte Institute of Political Studies, University of Tartu, Tartu, Estonia.
[26]Department of Nordic Studies and Linguistics, University of Copenhagen, Copenhagen, Denmark.
[27]National Library of Norway, Oslo, Norway.
[28]Helsinki Institute for Social Sciences and Humanities, University of Helsinki, Helsinki, Finland.
[29]Austrian Centre for Digital Humanities and Cultural Heritage, Austrian Academy of Sciences, Vienna, Austria.
[30]Department of Statistics, Uppsala University, Uppsala, Sweden.

*Corresponding author(s). E-mail(s): tomaz.erjavec@ijs.si; kopp@ufal.mff.cuni.cz; petya@bultreebank.org;
Contributing authors: nikola.ljubesic@ijs.si; taja.kuzman@ijs.si; p.rayson@lancaster.ac.uk; maciej.ogrodniczuk@ipipan.waw.pl; ccoltekin@sfs.uni-tuebingen.de; danijel@pja.edu.pl; katja.meden@ijs.si;

jure.skubic@inz.si; peter.rupnik@ijs.si; agnoloni@igsg.cnr.it; jagc@edu.ulisboa.pt; starkadur.barkarson@arnastofnun.is; roberto.bartolini@ilc.cnr.it; nuria.bel@upf.edu; calzada@uji.es; roberts.dargis@lumii.lv; sascha.diwersy@univ-montp3.fr; maria@athenarc.gr; r.j.vanheusden@uva.nl; mikel.iruskieta@ehu.eus; neeme.kahusk@gmail.com; ganna.kryvenko@inz.si; ligeti-nagy.noemi@nytud.hun-ren.hu; mariadelcarmen.magarinos@usc.gal; martin.molder@ut.ee; costanza@hum.ku.dk; kivs@bultreebank.org; lars.tungland@nb.no; jouni.tuominen@helsinki.fi; j.vidler@lancaster.ac.uk; adina.vladu@usc.gal; tanja.wissik@oeaw.ac.at; vaino.yrjanainen@statistik.uu.se; darja.fiser@inz.si;

Abstract

The paper presents the results of the ParlaMint II project, which comprise comparable corpora of parliamentary debates of 29 European countries and autonomous regions, covering at least the period from 2015 to 2022, and containing over 1 billion words. The corpora are uniformly encoded, contain rich metadata about their 24 thousand speakers, and are linguistically annotated up to the level of Universal Dependencies syntax and named entities. The paper focuses on the enhancement made since the ParlaMint I project and presents the compilation of the corpora, including the encoding infrastructure, use of GitHub, the production of individual corpora, the common pipeline for producing their distribution, and use of CLARIN services for dissemination. It then gives a quantitative overview of the produced corpora, followed by the qualitative additions made within the ParlaMint II project, namely metadata localisation, the addition of new metadata, such as the political orientation of political parties, the machine translation of the corpora to English and its tagging with semantic classes, and the production of pilot speech corpora. Finally, outreach activities and further work are discussed.

Keywords: Parliamentary proceedings, Comparable corpora, TEI

1 Introduction

Parliamentary proceedings, i.e. transcripts of debates in the highest democratic body of a country or autonomous region, have two characteristics that make them an especially good text type to compile into language corpora. Given the huge impact of their content, they are, on the one hand, of interest to a wide spectrum of researchers from political science, history, sociology, linguistics, discourse analysis, sociolinguistics, as well as citizen science. On the other hand, the transcripts are very easy to obtain directly from the internet, and have, unlike most other corpora, no copyright, privacy protection or terms-of-use barriers to their collection, processing and dissemination. It is therefore not surprising that many corpora of parliamentary proceedings have already been compiled (Fišer & Lenardič, 2018; Lenardič & Fišer, 2023), and there are numerous studies of parliamentary speeches that explored various themes, e.g. a study on populism and the strategies employed by the MPs in representing and involving people in parliamentary discourse (Truan, 2019), a discourse analysis providing insights into the treatment of female politicians (Stopfner, 2018) or a study on representation of what is deemed "uncivilised" (people, places and practises) across the past two centuries (Alexander & Struan, 2022).

However, as a rule, the existing corpora cover a single parliament, with, so far, almost no attempts (but see Truan and Romary (2022) and Sylvester, Greene, and Ebing (2022) for two exceptions) to develop a large and comparable set of corpora of national parliamentary proceedings.

The ParlaMint I project (2020–2021) produced a set of comparable parliamentary corpora of 17 European national parliaments with almost half a billion words, mostly starting in or before 2015 and ending in mid-2020, with the corpora uniformly encoded and containing rich metadata about the 11 thousand speakers. In addition to this "plain text" set of corpora, a linguistically annotated version was also released and both were made openly available for download and analysis through concordancers (Erjavec, Ogrodniczuk, et al., 2023).

This paper presents the results of the continuation of the project, ParlaMint II (2022–2023), which enlarged the set of corpora to 29 European countries and autonomous regions (c.f. Figure 1), extended the time coverage to at least 2022, and introduced other enhancements. In the scope of ParlaMint II, three versions of the corpora were published: 3.0, an intermediate project release, 4.0, the final project release[1] and 4.1 as a maintenance release completed after the project's end, which corrects some errors found in 4.0 and extends the time-frame of the UA corpus. In this paper, we present version 4.1.[2]

The paper focuses on the enhancement introduced in ParlaMint II and is structured as follows: Section 2 describes the compilation of the corpora, including the encoding, use of GitHub, a short per-corpus overview, the common pipeline for finalising the corpora, and the use of CLARIN services for dissemination; Section 3 gives a quantitative overview of the produced corpora, i.e. basic statistics of the corpora, of the speakers and their affiliations, and of the speeches; Section 4 discusses the qualitative additions made within the ParlaMint II project, namely the metadata localisation, the addition of new metadata, the machine translation of the corpora to English and its semantic tagging, and the production of pilot speech corpora; and Section 5 gives the conclusions, including outreach activities and a discussion of plans for further work.

2 Corpus compilation

Both in ParlaMint I and ParlaMint II, the individual partners were responsible for producing ParlaMint-compatible corpora of their parliament rather than these being

[1] The 4.0 release comprises the "plain text" corpora (Erjavec, Kopp, Ogrodniczuk, Osenova, Agirrezabal, et al., 2023), the linguistically annotated corpora (Erjavec, Kopp, Ogrodniczuk, Osenova, Agerri, et al., 2023), and the corpora machine-translated to English (Kuzman et al., 2023).

[2] The ParlaMint encoded 4.1 is ready but the full corpus build and service integration takes several weeks, with possibly more than one iteration. Version 4.1 will be released in time for the revised paper, if accepted.

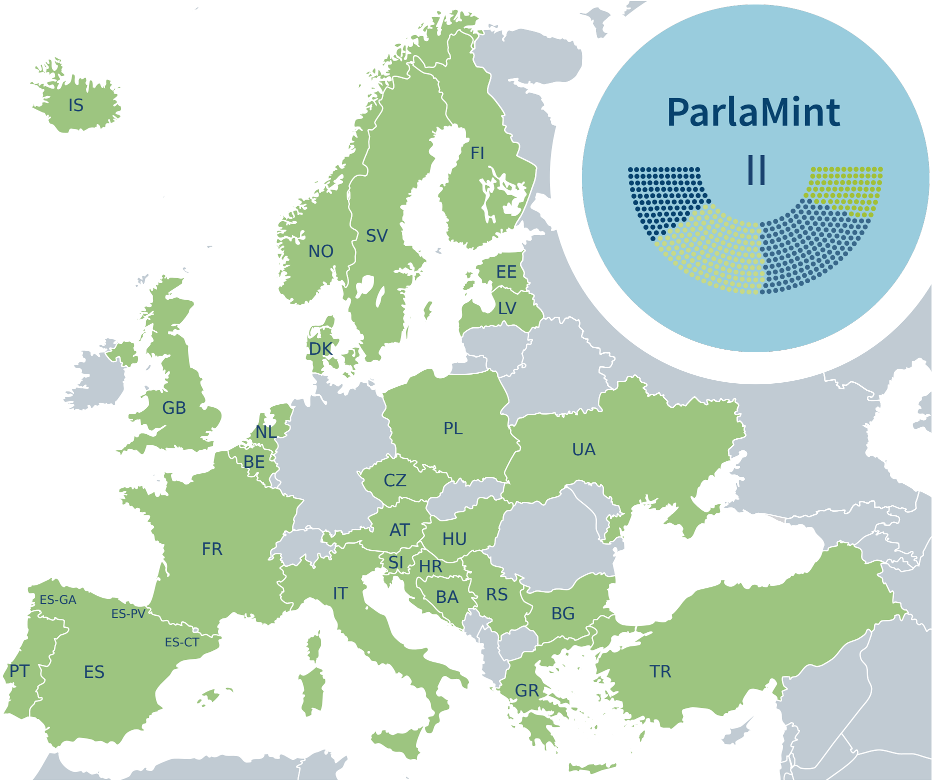

Figure 1 Coverage of ParlaMint corpora. The codes for countries and autonomous regions follow ISO 3166 "Codes for the representation of names of countries and their subdivisions" and are used in the rest of the paper.

centrally gathered and compiled. It was therefore important to ensure good annotation guidelines, a robust and versatile collaborative environment and validation procedures, to prevent errors and facilitate interoperability of the released set of corpora. In this section, we overview these aspects of the project, as well as giving a short overview of the related work on the individual ParlaMint corpora. Furthermore, we also explain the workflow for finalising the corpora, and their distribution via the CLARIN infrastructure.

2.1 The ParlaMint encoding

The XML schema used for validation and schema-aware editing of ParlaMint I corpora was based on the Parla-CLARIN recommendations[3] (Erjavec & Pančur, 2022), a customisation of the Text Encoding Initiative (TEI) Guidelines[4] (TEI Consortium, 2017). However, the ParlaMint I schema was written directly in the XML validation

[3] https://github.com/clarin-eric/parla-clarin/

[4] https://tei-c.org/

language RelaxNG, rather than being defined by a TEI ODD document, which customises TEI for a particular project or purpose. An ODD (One Document Does it all) should also contain the prose annotation guidelines, while the element and attribute specifications are accompanied by explanatory prose and examples.

In ParlaMint II, we first revised the Parla-CLARIN recommendations to take into account lessons learned in the ParlaMint I project, while still maintaining their broad applicability. Next, we wrote the ParlaMint ODD, based on the Parla-CLARIN one, which further constrains the schema specification and gives detailed ParlaMint-related prose guidelines. In the schema specification, we also substituted many generic TEI descriptions of elements and examples of their use with ParlaMint-specific explanations and snippets from the actual corpora. The ParlaMint (and, where relevant, Parla-CLARIN) recommendations were also extended with the new types of annotations introduced in ParlaMint II (cf. Section 4).

As with Parla-CLARIN, the ParlaMint TEI ODD schema is compiled into a RelaxNG schema for XML validation or other processing, such as XML schema-aware editing, while the guidelines, as well as the schema specification are compiled into HTML for reading. At the same time, we are still using (and updating) the ParlaMint I type RelaxNG schemas, as they have the advantage of quick fixes, more fine-grained control, and allow for validation of separate files not rooted in the `<TEI>` or `<teiCorpus>` elements, a requirement of ODD-derived validation. This means a certain amount of overhead, as each change has to be implemented and each document validated twice, however, it offers greater flexibility in developing and using the ParlaMint schemas.

ParlaMint I also established precise rules for the naming and structure of files and directories of a corpus, and these have not changed in ParlaMint II. However, there was one change that impacted the number of files. In ParlaMint I, a corpus root file contained the complete corpus TEI header (and XIncludes of the corpus components, i.e. transcriptions), which includes taxonomies (controlled vocabularies), and the list of speakers and of organisations. The latter two made the central file of a corpus very large, and so unwieldy (in editors) or impossible (in GitHub) to display, complicating its maintenance. Furthermore, ParlaMint II made a concerted effort to unify and localise (translate) its taxonomies into the ParlaMint II languages (cf. Section 4) and having taxonomies as part of each root file also complicated this development.

For these reasons, we factored out the files for the speakers, organisations, and the eight ParlaMint II taxonomies, with the files XIncluded in the TEI header of the corpus root file. Note that specifics of particular parliaments could still be expressed in local taxonomies, in which case the corpus includes two types of taxonomies for the relevant metadata dimension: the common and the corpus-specific one.

In ParlaMint I, there were taxonomies for legislature, speaker types, and subcorpora, and in the linguistically analysed version, also for Universal Dependencies (UD) (de Marneffe, Manning, Nivre, & Zeman, 2021) syntactic labels and the standard 4-class named entities (Tjong Kim Sang & De Meulder, 2003).

In ParlaMint II, we unified the UD labels by automatically deriving the taxonomy (i.e. the list) of labels with their glosses from the UD GitHub repository.[5] We also

[5] https://github.com/UniversalDependencies/docs/tree/0749864b5048bb8995fe68aedc37f721bc1338ee

added two taxonomies for political orientation, and one for USAS semantic classes (cf. Section 4).

2.2 Use of GitHub

GitHub was already used in ParlaMint I, where it not only supported revision control of all ParlaMint schemas and tools[6] but was also central to setting up the corpus compilation workflow. In ParlaMint II, due to the much larger number of partners, all detailed technical discussions were moved to GitHub issues,[7] while the aforementioned ParlaMint encoding guidelines were made available on the GitHub pages.[8] This step also had a significant impact on the corpora already included in the ParlaMint I project, as not only were they expanded in ParlaMint II, but more rigorous validation procedures (including manual corpus verification by the corpus editors) were applied, discovering various errors or potential changes needed to make the corpora even more consistent and interoperable. A number of such issues are still open, but they represent valuable (and public) documentation about the problems that have already been discovered.

The workflow for producing the individual corpora is based on the idea that a contributor of a corpus forks the main ParlaMint repository on GitHub, inserts a sample of their ParlaMint corpus to the fork, and then makes a pull request once the sample is compatible with ParlaMint. Ideally, this involves using the supplied and self-documenting `Makefile` to validate their sample and down-convert it to other formats,[9] with the partner then checking them for errors. Even if local validation is not possible (e.g. due to lack of access or lack of familiarity with Linux), a pull request to the repository triggers validation and down-conversion using GitHub Actions.

Following the partner's submission of a pull request and successful or almost successful automatic validation, a corpus editor verifies the sample. Subsequently, an issue containing a list of identified errors or suggested improvements for the sample is created. This issue is then used to discuss specific problems related to the sample. Once the issues are resolved, the sample is merged into the main repository in all formats. The sample can then be cited and commented on in issues, used in the documentation, or used directly as an example for other compilers of a ParlaMint corpus. Once valid samples are available, the partners would move on to producing the complete corpus, which would be collected and processed centrally (including validation, cf. Section 2.4) to make a distribution.[10]

In practice, this workflow, together with on-going revisions of the encoding, was somewhat complicated to implement, mainly because the structuring of the samples was somewhat different from that of the complete corpora.[11] Nonetheless, despite the complications, Git and GitHub were generally accepted by the ParlaMint partners.

[6] ParlaMint toolbox is written in XSLT and Perl, and the whole environment depends only on software found on Linux systems, as well as some easily obtainable support tools.

[7] To date, over 450 issues have been posted, many with detailed discussions.

[8] https://clarin-eric.github.io/ParlaMint/

[9] The down-conversion itself also uncovers errors, as scripts may issue error messages or fail to complete, and the generated CoNLL-U files are validated with the official Universal Dependencies validator.

[10] While it would be ideal to store complete corpora in Git, the number and total size of files make this difficult.

[11] This has been now simplified, partly due to the common pipeline discussed in Section 2.4

Erjavec, Kopp, and Meden (2024) present a survey among the partners about their experiences with Git(Hub). The survey collected 35 responses and the answers show a generally positive experience with the communication and workflow throughout the process, although not everyone was very happy with the use of GitHub issues and most complained about the differences between the production of the samples and the complete corpora. The group of (digital) humanities participants, as expected, generally had more difficulties with Git(Hub) and the workflow compared to the group of non-DH participants, which consisted mainly of computer scientists and/or computational linguists. However, both groups agreed that they are very likely to use Git in their future work.

2.3 Compiling individual corpora

As mentioned, the partners produced their ParlaMint-encoded corpora individually, as well as performed their linguistic analysis and mark-up. For ParlaMint I, the corpus compilation of the individual corpora was described in Erjavec, Ogrodniczuk, et al. (2023), while the number of partners precludes such a comprehensive description for ParlaMint II. However, this information is, in ParlaMint II, readily available in the `README` files for each corpus in the `ParlaMint/Samples/` directory on GitHub.[12] Nevertheless, to present the related work on the individual corpora, we here give, first, a list of those corpora that have publications on how they were compiled, and, second, a Table enumerating the tools that were used for the linguistic annotation.

The following corpora have published work on their compilation:

AT: The ParlaMint corpus is based on the ParlAT Corpus (Wissik & Pirker, 2018), which had a slightly different encoding (Wissik, 2022) from the ParlaMint one.

CZ: The source for the ParlaMint corpus was the Czech parliamentary corpus ParCzech 4.0 (Kopp, 2024b), which has slightly extended the ParlaMint schema in order to have more detailed named entities and audio alignment. The development process of a previous version of this corpus is described in Hladká, Kopp, and Straňák (2020); Kopp, Stankov, Krůza, Straňák, and Bojar (2021).

IS: The compilation of a previous version of the corpus is described in Steingrímsson, Barkarson, and Örnólfsson (2020).

IT: A detailed description of a previous version of the corpus is given in Agnoloni et al. (2022).

SI: The source for the ParlaMint corpus was siParl 3.0 (Pančur et al., 2022), with a previous version of siParl described in Pančur and Erjavec (2020).

UA: The corpus corpus compilation method is described in Kryvenko and Kopp (2023).

Once the plain-text version of each corpus was ready, it had to be linguistically annotated. It was up to the partners which tools to use for this task, and Table 1 presents their overview.

It can be seen that numerous tools were used for linguistic annotation, however, with certain (multilingual) tools being employed for a number of corpora. In particular, UDPipe was used for eight corpora, CLASSLA-Stanza for five, Stanza for four, and NameTag also for four corpora.

[12] https://github.com/clarin-eric/ParlaMint/tree/main/Samples

Table 1 Overview of tools used to linguistically annotate the individual ParlaMint corpora for their four annotation layers: segmentation into tokens and sentences ¨, morphological analysis and lemmatisation ●, syntactic analysis ▴, and Named Entity Recognition ○.

ID	Linguistic annotation
AT	¨●▴ UDPipe (Straka, 2018), ○ NameTag (Straková, Straka, & Hajič, 2019)
BA	¨●▴○ CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)
BE	¨● int-tagger, ▴ UDify (Kondratyuk & Straka, 2019), ○ flair-ner (Akbik et al., 2019)
BG	¨●▴○ CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)
CZ	¨●▴ UDPipe (Straka, 2018), ○ NameTag (Straková et al., 2019)
DK	● cstlemma (Jongejan & Dalianis, 2009), ¨●▴ UDPipe (Straka, 2018), ○ CST-NER
EE	¨○ EstNLTK (Laur, Orasmaa, Särg, & Tammo, 2020), ●▴ Stanza (Qi, Zhang, Zhang, Bolton, & Manning, 2020)
ES-CT	¨●○ Freeling, ▴ UDPipe (Straka, 2018)
ES-GA	¨● Freeling, ▴ UDPipe (Straka, 2018), ○ NER
ES-PV	¨●▴ UDPipe (Straka, 2018), ○ XLM-RoBERTa
ES	¨●▴ UDPipe (Straka, 2018), ○ NameTag (Straková et al., 2019)
FI	¨●▴ NLP-pipeline (Tamper, Leskinen, Apajalahti, & Hyvönen, 2018), ○ Nelli-Tagger (Tamper, Oksanen, Tuominen, Hietanen, & Hyvönen, 2020)
FR	¨●▴○ Stanza (Qi et al., 2020)
GB	¨●▴○ stanford-corenlp (Manning et al., 2014)
GR	¨●▴○ ILSP Neural NLP Toolkit for Greek (Prokopidis & Piperidis, 2020)
HR	¨●▴○ CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)
HU	¨●▴○ huspacy (Orosz, Szántó, Berkecz, Szabó, & Farkas, 2022)
IS	¨ tokenizer, ● abltagger-pos, ● nefnir, ○ IcelandicNER (Guðjónsson, Loftsson, & Daðason, 2021), ●▴ combo-ud (Jasonarson, Steingrímsson, Sigurðsson, & Daðason, 2022)
IT	¨●▴○ Stanza (Qi et al., 2020)
LV	¨●▴○ LV-NLP-PIPE (Znotins & Cirule, 2018)
NL	¨● int-tagger, ▴ udify (Kondratyuk & Straka, 2019), ○ flair-ner (Akbik et al., 2019)
NO	¨●▴○ Spacy (Honnibal, Montani, Van Landeghem, & Boyd, 2020)
PL	● app-morfeusz, ● app-concraft, ○ app-liner, ●▴ app-combo
PT	¨ LX-tokenizer (Branco & Silva, 2004), ●○ MBT-tagger, ▴ LX-UD (Branco, Silva, Gomes, & António Rodrigues, 2022)
RS	¨●▴○ CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)
SE	¨●▴○ Stanza (Qi et al., 2020)
SI	¨●▴○ CLASSLA-Stanza (Ljubešić & Dobrovoljc, 2019; Terčon & Ljubešić, 2023)
TR	¨● TRmorph (Ç. Çöltekin, 2010), ▴ steps-parser (Grünewald, Friedrich, & Kuhn, 2021), ○ TurkishNER
UA	¨●▴ UDPipe (Straka, 2018), ○ NameTag (Straková et al., 2019)

In addition to the linguistic analysis as such, each partner also had to convert their ParlaMint-encoded corpus into the format that could serve as the input to the linguistic annotation tool, and then insert the linguistic annotations into their corpus. Here, the biggest challenge turned out to be dealing with the transcribers' comments which were located directly inside paragraphs, i.e. mixed with the annotated text. Some XML tools for this merging, in particular those in the pipeline used to make the ParCzech corpus (Kopp, 2022) and that used for BA, HR, SR, and SI corpora were used in the context of cross-team assistance for other corpora as well.

2.4 The pipeline for corpus distribution

While each partner produced their ParlaMint corpus, there was nevertheless some central processing to compile the corpora to the datasets that form a part of a distribution.

First, the newly localised metadata and the added TSV-formatted metadata (cf. Section 4) were added to the corpora. Second, there were certain details that were observed to be wrong in the submitted corpora, and while each partner was notified of the problems, not everybody was able to correct them (e.g. because the person who produced the corpus was no longer available), so a script was written that corrected those errors that could be fixed automatically (while the others were reported in GitHub issues). Third, the TEI header of each corpus contains a fair amount of redundant (as it can be computed) metadata on the corpora, such as extents, quantitative information about the usage of tags, boilerplate titles, the version of the corpus etc., and the third script adds this metadata to the corpora in case it had not been inserted already or was wrong. The ParlaMint-wide taxonomies are also reduced to English and the language of the corpus and stored together with the corpus. With these three processing steps, the final ParlaMint-encoded corpora for a particular release have been compiled.

The next stage involves producing the corpora as they are present in the distribution. Extensive validation is performed first, not only via the ParlaMint RelaxNG and ODD schemas, but also checking the validity of all links, and, with a dedicated XSLT script, content validation which cannot be performed with XML schemas. The script produces extensive log files with informative, warning, and error messages. After validation, down-conversions are performed, which transform the corpus into simpler and directly usable formats, i.e. plain text, CoNLL-U files, per-speech TSV metadata files, as well as vertical files for the concordancers. The last operation is packaging the corpora in all the formats (and adding `README`s) as `.tgz` files for uploading to the repository.

2.5 Use of CLARIN services

As with ParlaMint I, the complete corpora are available for open (CC BY) download from the CLARIN.SI repository of language resources and tools.[13] In addition to the corpora, each repository entry also contains the log files produced by the corpus compilation pipeline, as well as the GitHub files corresponding to the release.

The corpora are also available for on-line exploration. As in ParlaMint I, they are mounted on the CLARIN.SI concordancers, in particular the noSketch Engine[14] (Kilgarriff et al., 2014) and KonText[15] (Machálek, 2020).

A new addition in ParlaMint II is the integration of corpora into the TEITOK web-based corpus platform[16] (Janssen, 2016). This platform not only enables users to query the corpus but also broadens access to parliamentary data for a diverse audience through the incorporation of a browsing feature. This feature facilitates the reading

[13] https://www.clarin.si/repository/xmlui

[14] Without log-in (https://www.clarin.si/ske) and with log-in (https://www.clarin.si/skelog), which provides more functions.

[15] https://www.clarin.si/kontext

[16] https://lindat.mff.cuni.cz/services/teitok/parlamint-40/

of transcripts and allows users to seamlessly switch between multiple view modes, enabling them to select the mode that best aligns with the specific demands of their research domain. Additionally, user can also explore persons, organisations and their relations.

3 Overview of the corpora

This section gives quantitative information about the current version of the ParlaMint corpora, in particular some basic statistics in terms of the languages used, their time span and size, statistics over the main metadata about the speakers, and over the speeches, i.e. transcriptions.

3.1 Basic statistics

ParlaMint version 4.1 comprises 29 corpora with 30 main languages[17] containing 8 million speeches and 1.2 billion words. Table 2 gives a quantitative overview of some basic characteristics of the individual corpora.

The first column gives the country codes of the corpora, and the second column the ISO 639-1 code of the main language(s) used in the corpus. Language is identified on the paragraph (technically, the `<seg>` element) level which appears inside speeches, as some speakers switch between languages.[18] Out of the 29 corpora, 6 are bilingual, and the table gives the predominant language first. It should be noted that some corpora mark snippets (individual speeches or paragraphs) in other languages, in particular English and French.

The third column contains labels for parliamentary bodies included in the transcripts: unicameral parliament, lower and/or upper house for bicameral parliaments, and parliamentary committees. This is important information for the comparability of the corpora, as it is sensible to compare the speeches of the same type of body, although most likely treating unicameral parliaments and lower house as the same type. Most corpora also contain these two sets of transcriptions, which some (Great Britain, Netherlands, Poland) contain transcripts of both the upper and lower house. The Norwegian corpus contains labels for both unicameral, as well as for lower and upper houses because in 2009 Norway changed its parliamentary system from a (pseudo-)bicameral to a unicameral one. The only corpus containing only the transcripts of the upper house is the Italian one. The Belgian corpus is currently the only one in ParlaMint that also includes the sessions of various parliamentary committees.

The next three columns give time-related information on the corpora, starting with the number of (possibly partial) terms[19] that the corpus covers. These largely reflect the time-frame of the corpus, but also indicate the dynamics of (possibly extraordinary) elections. The From and To dates and, hence, the number of years of included speeches vary considerably, with almost all starting in or before 2015 and ending in 2022. The only corpus that starts after 2015 is the French one (starting mid 2017, and,

[17] Or 29, if the NO language varieties Bokmål and Nynorsk are taken as one language, i.e. Norwegian.

[18] UA additionally identifies the language on the sentence level. The paragraph language is set to the language that has more tokens in paragraph.

[19] The number of terms (elections) refers to those of the lower house, if it is present in the corpus, of the upper house for the rest.

Table 2 Basic information about the ParlaMint corpora including the corpus country or region code (ID), the language(s) of the corpus (Lang), the parliamentary bodies included (Bodies = uni / unicameral parliament, upp / upper house, low / lower house, com / parliamentary committees), the number of terms included in the corpus (Ts), start (From) and end (To) month of included transcripts, the number of years covered (Yr), the number of millions of words per year (Mw/Yr) and in total (Mw).

ID	Lang	Bodies	Ts	From	To	Yrs	Mw/Yr	Mw
AT	de	low	8	1996-01	2022-10	27.1	2.24	60.84
BA	bs	uni	7	1998-11	2022-07	24.0	0.76	18.31
BE	fr+nl	low+com	2	2014-06	2022-07	8.2	5.42	44.37
BG	bg	uni	5	2014-10	2022-07	7.9	3.37	26.47
CZ	cs	low	3	2013-11	2023-07	9.8	3.14	30.77
DK	da	uni	4	2014-10	2022-06	7.8	5.25	40.80
EE	et	uni	3	2011-04	2022-06	11.4	2.01	22.87
ES-CT	es+ca	uni	4	2015-10	2022-07	6.8	2.33	15.95
ES-GA	gl	uni	3	2015-01	2022-05	7.4	2.40	17.84
ES-PV	eu+es	uni	3	2015-02	2022-07	7.5	1.80	13.54
ES	es	low	5	2015-01	2023-02	8.2	2.39	19.65
FI	fi+sv	uni	2	2015-04	2022-01	6.9	1.98	13.54
FR	fr	low	2	2017-06	2022-03	4.8	10.33	49.63
GB	en	low+upp	4	2015-01	2022-07	7.6	16.56	126.71
GR	el	uni	3	2015-01	2022-02	7.2	6.91	49.70
HR	hr	uni	5	2003-12	2022-07	18.8	4.64	87.32
HU	hu	uni	3	2014-05	2023-07	9.4	3.29	30.85
IS	is	uni	4	2015-01	2022-07	7.6	4.10	31.19
IT	it	upp	2	2013-03	2022-09	9.7	3.31	31.97
LV	lv	uni	2	2014-11	2022-10	8.1	1.13	9.16
NL	nl	low+upp	5	2014-04	2022-09	8.5	7.86	66.85
NO	nb+nn	uni+low+upp	7	1998-10	2022-09	24.3	3.63	88.45
PL	pl	low+upp	4	2015-11	2022-06	6.7	5.35	36.06
PT	pt	uni	3	2015-01	2022-03	7.3	2.41	17.65
RS	sr	uni	9	1997-12	2022-07	25.0	3.38	84.57
SE	sv	uni	2	2015-09	2022-05	6.8	4.28	28.98
SI	sl	low	6	2000-10	2022-05	21.9	3.20	69.92
TR	tr	uni	4	2011-06	2022-11	11.6	4.26	49.26
UA	uk+ru	uni	6	2002-05	2023-11	21.8	1.93	42.00

as the shortest corpus, containing less than 5 years, ending in 2022), while many others start much sooner, with the Austrian one going as far back as 1996, and covering, as the longest corpus, over 27 years. As for the end dates, the Finish corpus ends in January 2022, while, on the other hand, the Czech and Hungarian one extend to July 2023, and the Ukrainian one all the way to November 2023.

Finally, the last two columns give the size of each corpus in words per year and as a whole. By far the largest corpus, both per year and in total, is that of Great Britain (16 and 126 million), 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, which must be a result of longer or more sessions of their parliaments. In the opposite direction, the outliers are the Bosnian corpus (only .76 million words per year) and the Latvian corpus (only 9 million words in total). The former has relatively few sessions, while the latter covers less years than the others, except for France.

3.2 Metadata on speakers

The ParlaMint corpora contain significant metadata about its 24,021 speakers, which allows for various political or sociological but also linguistic studies for which speaker-related variables are required. Table 3 gives an overview of speaker-related data over the individual corpora.

Table 3 Metadata on speakers divided into three groups. The first relates to (political) organisations, the second to persons, and the third to their affiliations to organisations. The first group consists of the number of defined organisations (Org), political parties and parliamentary groups (Prt), coalitions and oppositions (C/O); the second of the number of defined persons (Pers), with known sex (Sex), birth date (Birth), and with Web link(s) (URL). The third group gives the number of defined affiliations (Affil), the number of ministers (Mini), members of parliament (MPs), and members of political parties or parliamentary groups (PrtyM).

	Organisation			Person				Affiliation			
ID	Org	Prt	C/O	Pers	Sex	Birth	URL	Affil	Mini	MPs	PrtyM
AT	37	18	38	854	854	848	854	3,456	122	776	795
BA	42	40	14	603	282	231	0	823	24	282	278
BE	68	66	19	786	569	569	0	2,174	35	551	551
BG	46	38	6	1,032	1,032	912	95	4,559	25	838	817
CZ	450	33	7	597	570	507	572	13,632	93	536	465
DK	21	19	8	383	383	383	0	1,025	73	383	383
EE	8	6	7	488	264	263	0	1,083	62	263	262
ES-CT	39	37	5	364	364	364	0	1,726	44	324	364
ES-GA	59	57	6	227	227	214	182	722	16	170	212
ES-PV	11	9	5	197	197	175	156	440	21	193	193
ES	52	50	10	941	926	884	0	1,849	65	843	826
FI	19	17	16	314	310	310	0	1,187	77	306	305
FR	185	26	5	908	908	902	0	2,622	18	846	814
GB	37	34	5	1,951	1,951	0	1,951	9,120	80	1,868	1,947
GR	16	14	5	635	635	0	0	2,562	91	532	532
HR	47	45	12	1,036	660	660	0	2,373	78	660	660
HU	94	38	6	492	492	488	0	3,420	25	279	343
IS	12	9	5	261	261	261	1	925	26	138	239
IT	47	45	23	771	771	771	771	3,249	82	706	597
LV	13	11	6	234	234	0	0	488	35	196	196
NL	50	35	14	586	586	542	557	1,140	49	244	549
NO	17	13	9	1,106	1,106	1,106	0	5,067	141	1,069	1,106
PL	12	9	4	1,223	1,223	753	753	2,180	53	753	645
PT	25	22	8	723	723	665	0	2,591	52	601	709
RS	73	71	18	1,724	1,472	1,472	0	4,992	57	1,472	1,472
SE	15	13	5	649	649	0	0	1,947	49	626	644
SI	32	29	30	973	973	466	330	1,646	59	415	410
TR	83	47	3	1,346	1,234	1,234	1,204	7,326	96	1,218	1,229
UA	151	148	38	2,617	2,617	2,459	528	10,661	225	1,827	1,826

The first group of the three columns relates to organisations. In the corpora, each organisation is given an ID, its full and abbreviated name, and, depending on the corpus, also the dates of its existence. The first numerical column gives the number of such entities, followed by political parties[20] only. It should be noted that corpora differ in terms of which organisations (as well as affiliations, i.e. the last group in the table) they encode: some encode only those that fall into the time-frame of their corpus, while others give the complete history of the persons and hence their organisations. The last column in this group gives the number of time-stamped coalitions and (for some corpora) oppositions of parliamentary groups (C/O).

The second group of four columns gives the numbers related to the defined persons. The low numbers typically belong to regional parliaments (e.g. Basque country) or countries with a small population (Iceland), but are also dependent on the time span of the corpus, as a larger time-span will involve more speakers. The next three columns give the number of persons with additional personal details. The first is if they have a specified sex (useful for gender studies). All corpora have this information, if not for all speakers, then at least for the MPs. Next is the date of birth (for age-correlated studies), which is present in 25 corpora, with the last one whether they are associated with one or more URLs (Wikipedia page, official government Web page, Twitter or Facebook account), which could be of use for discovering more information about speakers, as well as for named entity linking; however, this information is available for only 12 corpora, and, except for AT, GB, and IT, for only an (often small) subset of the speakers.

The last group of four columns quantify the numbers related to affiliations of persons with organisations. The first column gives the number of affiliations that persons have, together almost 95 thousand affiliations or, on average, 4 affiliations per person. The minimum here is BA with 1.4 affiliations per person, while the maximum is CZ with 22.9, as it gives the complete affiliation history of a person; without CZ the average is 3.5. The affiliations also specify the role of the person in the organisation, as well as (for most corpora) the dates of the affiliation. The last three columns give the numbers of persons with particularly important affiliations: the first is the number of ministers, the second the number of MPs, and the third the number of people who are members of political parties or parliamentary groups.

3.3 Speeches and associated mark-up

The ParlaMint corpora contain almost 8 million speeches and 10 million elements with related information. The former is given in the first, and the latter in the second block of columns in Table 4.

The first group gives, first, the number of speeches per corpus, with the minimum for Basque country (40,000) and with, surprisingly, given its small date range, France having the most (over 700,000), meaning that their speeches are much shorter, most likely more dialogues, rather than monologues. The next column gives the number of speeches that are marked with their speaker, which is important for investigations that take into account the characteristics of the speakers. All the corpora give the speaker for the vast majority of the speeches, with the least by Sweden but even here less than 1.5% are missing. The next column shows the numbers of speeches spoken by non-chairs of the session (MPs, government members or guests), potentially an important

[20] In the corpora, we distinguish political parties from parliamentary groups, i.e. groups of parties forming a common list in the parliament. In Table 3, we count both as “Parties”.

Table 4 Overview of the speeches in the corpora. The first group gives the number of speeches (Speeches), how many speeches have a defined speaker (W.Spks), and how many are not spoken by the chair of the sessions (W.NCs). The second block gives other markup related to the speeches, i.e. the number of marked-up headings (Heads), notes (Notes), vocal, kinesic and other incidents (Incidents), and missing pieces of the transcriptions (Gaps).

	Speeches			Other mark-up			
ID	Speeches	W.Spks	W.NCs	Heads	Notes	Incidents	Gaps
AT	231,759	231,759	106,717	0	680,688	337,759	15,162
BA	126,252	126,030	67,754	0	126,326	3,483	3,679
BE	199,305	198,684	156,960	0	508,639	992	4,535
BG	210,018	208,565	107,315	0	0	51,652	0
CZ	196,185	196,185	91,320	0	243,108	33,355	1,086
DK	398,610	398,610	190,641	14,302	14,302	0	0
EE	227,872	227,872	130,934	0	233,814	0	0
ES-CT	50,824	50,824	27,031	283	67,172	21,099	127
ES-GA	83,078	83,078	38,090	0	91,441	58,417	0
ES-PV	39,148	39,148	18,014	0	47,882	0	0
ES	76,369	43,886	32,739	2,640	5,886	71,182	694
FI	146,858	146,858	116,755	6,806	10,635	42,150	0
FR	714,860	697,095	621,806	22,123	22,126	91,128	0
GB	670,912	667,916	654,567	31,215	191,793	0	0
GR	342,274	342,274	220,760	7,578	365,334	54,775	1,263
HR	504,338	497,137	257,753	0	498,874	29,145	51,084
HU	116,346	116,325	57,726	0	154,671	99,632	37
IS	95,286	95,286	92,578	0	137	49,810	0
IT	172,796	172,796	93,162	13,170	193,510	74,054	0
LV	162,782	162,782	80,747	0	163,720	0	0
NL	609,248	609,248	445,589	6,100	783,558	0	0
NO	398,809	396,858	275,017	20,123	1,565,683	0	0
PL	228,326	228,326	122,443	686	241,406	248,396	1,606
PT	170,937	170,937	118,370	1,430	34,745	62,704	0
RS	316,069	315,896	156,156	0	318,697	4,203	1,786
SE	84,662	83,436	84,662	15,819	370,551	9,656	0
SI	311,354	311,354	153,770	4,706	392,734	3,668	38,654
TR	681,052	681,052	486,410	0	109,555	114,378	0
UA	429,437	429,417	221,701	0	730,120	22,819	1,157

piece of data, as chairs speak a lot but mostly on procedural matters, so studies will likely filter out the speeches by chairs. For most corpora, the chairs give around half of all the speeches, with two exceptions. The SE corpus does not mark the role of the speaker, which is why the number of all speeches in the table equal to the number of the speeches by non-chairs, while IS has only about 7% of speeches given by chairs; this is a result of the source data on their parliamentary web site, which provides speeches by chairs only for their introductory speeches, but not the short speeches in the middle of the sessions, where they are mostly just giving the word to the next speaker and similar.

The second group of columns quantifies the other elements that appear in the corpus texts. Namely, the transcripts also contain session or agenda titles, names of speakers or chairs etc., which have been, to varying extents, preserved in about half of the corpora and marked up as headings. The transcripts also contain many transcriber notes, i.e. remarks about time, voting, interruptions, applause, or unintelligible speech. Such commentary was identified and marked up in several ways. The default was to mark it up as notes (possibly with a type specifying what kind), while the other option is to use more precise elements, the sum of which is shown in the "Incidents" column; these elements are `<vocal>` (non-lexical vocalised phenomena, e.g. exclamations from the auditorium), `<kinesic>` (non-vocalised communicative phenomena, e.g. applause) and `<incident>` (non-communicative phenomena, e.g. coughing), again, possibly using type attribute to categorise these elements further. As can be seen, the corpora are not uniform in the treatment of these elements, most use both, but seven just notes, and one only incidents (BG); obviously, more work would be necessary to harmonise this encoding.

The last column gives the number of identified gaps in the corpora, which correspond to pieces of missing transcriptions, which are mostly due the transcriber noting that they could not understand or hear the speaker (e.g. because the microphone was not turned on), or, in certain cases that a part of the transcription was omitted by the corpus compiler, e.g. the table of contents or other tables. The two are distinguished by the value (`inaudible` vs. `editorial`) of their `reason` attribute. It should be noted that the numbers in the Table are given from the "plain-text" version of the corpus. The linguistically annotated version should have the same numbers, except for gaps. Here, the annotation pipeline used for some corpora had problems with parsing very long sentences, which were therefore omitted from the corpus, and this was also marked up with the `<gap>` element.

4 ParlaMint II additions

In addition to improving the infrastructure of the project, increasing the number of corpora and extending them in time, ParlaMint II also introduced other additions to the corpora which we overview in this section.

4.1 Localisation

A perennial question with monolingual language resources is in which language the metadata of the resource should be in: either in English, to make it maximally useful in an international setting, or in the language of the resource, to enable researchers from the corresponding country (or region) to analyse the data in their native language, and to maintain language equality. The ideal, of course, is to have the metadata in both languages. Already in ParlaMint I, certain metadata (e.g. titles of sessions), was present in the main language of the parliament, as well as in English. In ParlaMint II, we made a concerted effort to improve the localisation of the metadata in several ways.

First, most ParlaMint taxonomies (legislature, speaker types, subcorpora and left-to-right political orientation) were translated to most of the 29 main ParlaMint languages, and are now maintained centrally. This avoids different naming of categories for different corpora and constitutes a highly multilingual resource which might be interesting for other purposes and researchers.

The second improvement was driven by the machine-translated corpus. As it is not very useful to have the transcriptions in English, but names of speakers and affiliated organisations in the Cyrillic or Greek alphabet, we added transliterated names to the corpora.[21]

The third improvement, enabled by the first two, was the localisation (or, rather, i18n of the scripts) of specific down-conversions of the corpora, in particular to the metadata TSV files, and to the vertical files. For the former, the corpus distributions now include the metadata files both in original language, as well as in English. For the latter, the individual ParlaMint corpora on the concordancers have their metadata in the original language, while the machine-translated corpus (cf. Section 4.3), as well as the aligned joint corpus of all the 29 corpora, have metadata in English.

4.2 Adding metadata

In ParlaMint II, we also added metadata on individuals and organisations that had been identified as potentially useful but were missing from the ParlaMint I corpora. In the corpora where this information was previously missing, we identified the ministers and added these time-stamped affiliations to the corresponding individuals (cf. the column Minister in Table 3 with 1,873 persons). Wikipedia, government websites, etc. served as sources of this information.

A much more difficult concept was the second addition (modelled as states of organisations), namely the political orientation of political parties (Erjavec, Meden, & Skubic, 2023). The first source for this addition was the Chapel Hill Expert Survey Europe (Jolly et al., 2022), in particular, the CHES[22] Trend File 1999–2019 and CHES 2019, which adds countries such as Norway (NO), Iceland (IS) and Turkey (TR). Together, these two CSV files contain 85 variables on a specific (political) position for each party identified and each year covered.

Although this dataset provides valuable expert data, it only partially covers the ParlaMint corpora: CHES does not include all ParlaMint countries and no autonomous regions, its time span is shorter, it does not include all political parties, and not all variables are available for all parties or years. In addition, the CHES dataset provides numerical values for its numerous variables.[23] However, we also wanted to use a simpler set of discrete categories for the political parties or their affiliated speakers, namely their political orientation on the left-to-right (L-R) axis. While this distinction is somewhat simplistic and only expresses the political orientation one-dimensionally, it is nevertheless widely used and can provide valuable insights. To obtain the categories, we used Wikipedia as a second source, which covers most parties and usually provides information on the party's L-R orientation in the infobox. Wikipedia distinguishes 13 positions along this axis (Far-left, Left to far-left, Centre-left to left etc.) and another 5 that fall outside this continuum (e.g. Big Tent, Pirate Party). Using this approach, we were able to assign the L-R orientation to 834 of the 1,073 parties and parliamentary groups, i.e. we achieved a coverage of 78%. To enable a wider coverage, we have also implemented the option for the encoders of the corpora to add the L-R information themselves, although only a few have made use of this, namely BE, PT and UA.

[21] Transliteration was done using Perl's `Lingua::Translit` module, choosing as the most useful (simple to input yet readable) `Streamlined System BUL` for BG, `DIN 1460 UKR` for UA, and `ISO 843` for GR.

[22] https://www.chesdata.eu/

[23] One of the CHES variables is `lrgen`, i.e., the general position of a party on the L-R axis.

From a technical perspective, the addition of the metadata was done centrally, using a method that was tested already in ParlaMint I (for markup of coalition/opposition information), namely that the metadata is not inserted directly into the ParlaMint-encoded (i.e. TEI/XML) files of the corpus, but indirectly via TSV files. The workflow for adding each additional metadata dimension consists of three steps. First, a script is written that converts the already existing metadata (if any) in a corpus into a TSV file and initialises the TSV file by writing the header line and e.g. one political party name per line. The encoder then imports the TSV file into their preferred spreadsheet editor, enters the required data and exports it as a TSV file. Adding the metadata to the TSV file can of course also be done automatically when the appropriate inputs are made, as was the case with the CHES orientations (although the country and party identifier mapping was done manually). The last step of the pipeline again consists of a script that checks the validity of the new TSV metadata[24] and merges it into the corresponding XML file of a corpus (either that for `<listPerson>` or for `<listOrg>`).

This approach allows the encoder of the additional metadata to focus on the information to be entered rather than the intricacies of its XML encoding, and may also be useful in the future for adding further metadata that can be easily expressed in a tabular format. Kryvenko and Kopp (2023) highlight the significant benefits of this approach for UA corpus development, with the most important advantages being the facilitation of collaboration between humanities scholars and computer scientists and a clear distinction between automatic and manual data entry.

4.3 Machine translation and semantic annotation

To further benefit from the comparability and interoperability of the corpora and provide researchers with a possibility for investigating parliamentary phenomena across all ParlaMint corpora, the ParlaMint II project included the machine translation of the corpora into English, as well as semantic tagging of the translated corpora.

Machine translation[25] (MT) was performed with the pre-trained Transformer-based OPUS-MT models (Tiedemann & Thottingal, 2020). These models are built upon the MarianNMT neural machine translation toolbox (Junczys-Dowmunt et al., 2018) and were trained on parallel corpora from the OPUS repository (Tiedemann, 2012). The OPUS-MT models are either specialised for a specific language, such as models for Polish, or for a language family, such as models for South Slavic languages. The models for a language group were especially useful for cases where the corpus comprised debates in multiple related languages, such as Ukrainian and Russian in UA, or Catalan and Spanish in ES-CT. In contrast, if the corpora consisted of non-related languages, such as Dutch and French in BE or Spanish and Basque in ES-PV, they had to be split into two parts and processed separately. Prior to machine translation of the full corpora, a manual evaluation of samples machine-translated with all the available models was performed by the partners for each of 30 languages to determine which model provides the best results for each language.

The pipeline to produce the machine-translated English ParlaMint corpora involves several steps. First, the speeches are extracted from the CoNLL-U files, and

[24] Unfortunately, spreadsheet editors often silently change data and export it in a non-transparent way.
[25] The code for the MT pipeline is available at https://github.com/TajaKuzman/Parlamint-translation.

the transcriber notes from the ParlaMint-encoded files. Then, the notes and the sentences of the texts are machine-translated to English using the EasyNMT[26] library. To address the frequent inaccurately translated proper nouns, a post-processing step is performed by aligning[27] proper nouns with named entities extracted from the CoNLL-U files, using their lemmas as surface forms in the English translation. The translated corpora were then linguistically processed using the Stanza pipeline (Qi et al., 2020) on the same levels as the source-language corpora, except for syntax, which was too computationally demanding. For all levels, the default Stanza models were used which are trained on a combination of English Universal Dependencies datasets (Behzad & Zeldes, 2020; Monarch & Munro, 2021; Nivre et al., 2017; Silveira et al., 2014; Zeldes, 2017), except for the named entities for which we used the CoNLL03 model (Tjong Kim Sang & De Meulder, 2003) with 4 NER labels.

The already mentioned preliminary evaluation, despite being conducted on small samples, provided valuable insights into the translation quality. Overall, the machine translation output was found to be of high quality, however, approximately 20–30 % of the sentences still contained common machine translation errors. The errors can be on the word level, such as very frequent incorrect translations of proper nouns (e.g., *The Winner of the Welcomes* instead of *Zmago Jelinčič Plemeniti*, the name of a Slovenian politician) or incorrect translation of terms (e.g., *State Assembly* instead of *National Assembly*). Errors can also occur at the level of multi-word expressions (e.g., literal translation of *Besedo dajem* to *I give my word to* instead of *I give the floor to*), or at the utterance level, where we observed repetitions, additions, and hallucinations, that is, MT output that is not related to the source text. Therefore, it is crucial for any studies using translated corpora to clearly outline the limitations of using machine-translated content and to cross-check the findings with the source texts.

Semantic annotation of corpora can take multiple forms, including Word Sense Disambiguation (WSD) where an existing detailed ontology or taxonomy of fine-grained word senses is employed as a label set and one sense per word is assigned to each particular context using a variety of disambiguation methods to resolve ambiguity due to homonymy and/or polysemy. In general, semantic annotation can be useful for further tasks in an NLP pipeline or improving accuracy in applications such as information retrieval.

Our approach for ParlaMint II assigns coarse-grained semantic field labels from an existing tagset of 21 major top level domains (including 'emotion', 'money and commerce', and 'world and environment') at the top of a hierarchy splitting into 232 semantic tags.[28] The process used the UCREL Semantic Analysis System (USAS),[29] originally developed in C in the 1990s for English semantic annotation (Rayson, Archer, Piao, & McEnery, 2004) but recently released open source for multiple languages in the Python Multilingual UCREL Semantic Analysis System (PyMUSAS).[30]

[26] https://github.com/UKPLab/EasyNMT
[27] Word alignment was performed with https://github.com/robertostling/eflomal.
[28] https://ucrel.lancs.ac.uk/usas/USASSemanticTagset.pdf
[29] https://ucrel.lancs.ac.uk/usas/
[30] See https://pypi.org/project/pymusas/ and https://github.com/UCREL/pymusas

The English system relies on large manually created lexicons of single words and multiword expressions (MWEs),[31] and is around 91% accurate for English, annotating a variety of MWEs including phrasal verbs, noun phrases, proper names, named entities, multiword prepositions as well as non-compositional idiomatic expressions, which all receive one semantic tag across the whole MWE. Contextual disambiguation methods for the semantic tagger rely on a number of methods including part-of-speech tagging for filtering the range of semantic tags being considered, general likelihood ranking and heuristics for overlapping MWE resolution. On a practical level, the PyMUSAS pipeline includes a spaCy[32] PoS tagger, and for ParlaMint we applied it to the translated CoNLL-U files. The PyMUSAS annotation was highly parallelised on the Oracle Compute cloud taking approximately 12 hours for the whole corpus.[33]

As the final step in MT and semantic annotation, the original language ParlaMint-encoded corpora were first pre-processed to remove the content of all the sentences and transcriber notes, and to move the latter from inside the sentences to their beginning. Then the translated notes and additionally semantically annotated CoNLL-U files were ParlaMint-encoded and inserted into the pre-processed corpora, i.e. into the empty transcriber notes and sentences. These corpora were then finalised using the common pipeline for corpus compilation (cf. Section 2.4.) but with slight changes in the metadata, i.e. the language of the corpus, specifying that these are machine-translated corpora, and adding the taxonomy for the USAS semantic classes.

With this pipeline, the machine-translated and semantically annotated corpora are structured identically to the original ParlaMint corpora and also retain all their metadata. The resulting corpora are made available similarly to the original corpora, i.e. for download from the repository (Kuzman et al., 2023), and for analysis via the concordancers. For the concordancers, the ParlaMint corpora were joined into one corpus containing all the original language corpora, and one corpus containing all the machine-translated corpora, with both corpora constituting a parallel corpus aligned on the sentence level, and both with English language metadata.

4.4 Speech corpora

Speech corpora are typically expensive to construct and difficult to distribute as they have to be manually transcribed and contain biometric data. For ParlaMint corpora, neither applies: transcriptions are already available through the ParlaMint corpora, and, for many countries, the parliamentary audio/video is publicly available.

In ParlaMint II, we compiled pilot spoken corpora for four ParlaMint languages. Four datasets have been released so far, and they are detailed in Table 5.

The start of the Czech (speech) corpus construction (Kopp et al., 2021) predated ParlaMint II and was tailored to their specific data. On the other hand, the Croatian, Polish and Serbian corpora were compiled with a novel robust pipeline which can align a large collection of recordings with a large collection of transcripts, given no previous alignment, not even at the level of files.

[31] The lexicons for English and other languages are available for academic use with a Creative Commons licence, see https://github.com/UCREL/Multilingual-USAS

[32] https://spacy.io/

[33] We adapted a PyMUSAS CoNLL-U tagging script developed by Daisy Lal which is available at https://github.com/UCREL/pymusas-conllu-parlamint

Table 5 Currently available speech parliamentary corpora including the corpus country or region code (ID), name of the corpus, the number of hours of speech data, the number of sentences covered and the reference to the dataset.

ID	Corpus name	Hours	Sentences	Data
CZ	AudioPSP + ParCzech 4.0	4,590	1,976,928	(Kopp, 2024a, 2024b)
HR	ParlaSpeech-HR 2.0	3,061	922,679	(Ljubešić, Koržinek, & Rupnik, 2024)
PL	ParlaSpeech-PL 1.0	1,010	535,465	(Koržinek & Ljubešić, 2024)
RS	ParlaSpeech-RS 1.0	896	290,778	(Ljubešić, Rupnik, & Koržinek, 2024)

An early version of the alignment pipeline, along with the description of the Croatian ParlaSpeech-HR 1.0 corpus (Ljubešić et al., 2022) is described in Ljubešić, Koržinek, Rupnik, and Jazbec (2022). The alignment is complicated for several reasons: the transcripts do not have the same order as recordings, not all recordings are transcribed, nor all of those made public, and the transcripts sometimes follow the spoken word very vaguely (redaction, gaps, mistakes). To work around these issues, while scaling to thousands of hours of recordings and tens of millions of words of transcripts, our pipeline has the following steps. Voice activity detection is performed first and speech representations are extracted with a Transformer model. These representations are used to produce automatic transcripts. The ParlaMint transcripts are simplified and approximately matched to the generated transcripts. The best matching candidates are realigned on the word level with the help of speech representations. Finally, the word-level alignment is used to re-segment the matches to follow the ParlaMint transcript segmentation into speeches and segments.

The resulting ParlaSpeech corpora consist of audio segments that correspond to specific sentences in the transcripts. The transcripts contain word-level alignments to the recordings, allowing for simple further segmentation of long sentences into shorter segments for memory-sensitive applications. Each segment has a reference to the ParlaMint 4.0 corpus via utterance IDs and character offsets.

The speech corpora are not only available for download but also through the concordancers, where sentences are, for easier listening, further segmented into speech segments of up to 6 seconds around the concordance key. Finally, the corpora are made available through the HuggingFace Datasets,[34] allowing for simple usage of the data for fine-tuning Transformer models for automatic speech recognition or any other speech-related task.

5 Conclusions

The paper presented the current version of the ParlaMint corpora, including the infrastructure that enabled their compilation, and focusing on the additions achieved in the ParlaMint II project. Comprising 29 carefully structured corpora of parliamentary proceedings with over a billion words, significant metadata about the speakers, linguistic annotations, semantically-annotated machine translation to English, and featuring pilot speech corpora, the ParlaMint corpora should be a very valuable resource for anybody studying parliamentary discourse, especially in a comparative setting.

[34] The Croatian dataset can be accessed at https://huggingface.co/datasets/classla/ParlaSpeech-HR.

In addition to the presented work on the corpora, the ParlaMint II project also undertook dissemination activities. In 2023, two tutorials were given on the specifics and usage of ParlaMint corpora, one at the Digital Humanities conference in Graz (Kryvenko, Pahor de Maiti, & Osenova, 2023) and the other at the European Summer University in Digital Humanities (Kryvenko & Pahor de Maiti, 2023). The ParlaMint corpora were also used in tasks in the scope of two Helsinki Digital Humanities Hackathons. In 2022, a multi-disciplinary team investigated power distribution inside parliamentary networks using ParlaMint I corpora for GB, SI and ES, and with a special focus on gender distribution in the debates (Skubic, Angermeier, Bruncrona, Evkoski, & Leiminger, 2022). In 2023, using a draft edition of ParlaMint II corpora for GB, HU, SI and UA, the team investigated political polarisation, focusing on the topics of European Union, the war in Ukraine, and healthcare (Kryvenko, Evkoski, Bordon, & Meden, 2023). Furthermore, ParlaMint 4.0 will be used as the dataset in one of the themes of the 2024 edition of the hackathon with the title "Echoes of the Chambers: Studying Democracy through Parliamentary Speeches".[35] The ParlaMint II project was also presented to a large audience at the January 2024 "CLARIN Cafe".[36] Finally, a shared task using ParlaMint corpora with the title "Ideology and Power Identification in Parliamentary Debates" is to be held at CLEF 2024.[37] The sampled data for the task has been published (Ç. Çöltekin et al., 2024) with, currently, over 30 registered teams.

As regards further work, there would be a number of directions worth taking. First, it would be satisfying to fill in the grey areas presented in Figure 1, i.e. add the still missing European countries (and autonomous regions) to the ParlaMint set of corpora. Second, the current set of ParlaMint corpora mostly ends mid-2022, and it would be, of course, worthwhile to add the transcripts since then. For the new corpora, sites willing to get to grips with the ParlaMint encoding and compilation would need to be found, while for extending existing corpora, while the existing pipelines would most likely be able to handle the new transcripts, the addition of metadata (new terms, speakers and political parties) has to be, most likely, added manually. Third, ParlaMint has centrally produced the machine-translated and semantically-annotated versions, which would also need to be compiled for new or extended corpora. And fourth, ParlaMint has centrally added metadata, in particular the CHES datasets, currently reaching only 2019. The CHES datasets have just recently been updated,[38] and it would be beneficial to include this new information into ParlaMint, as well as extending such metadata with other sources, such as V-Dem[39] (Coppedge et al., 2020). It is the very richness of metadata and the annotations that makes the ParlaMint corpora difficult to maintain, and a synchronised effort to extend the ParlaMint corpora in number, time and metadata is most likely dependent on a new project that would support this effort.

[35] https://www.helsinki.fi/en/digital-humanities/dhh24-hackathon/dhh24-themes

[36] https://www.clarin.eu/event/2024/clarin-cafe-parlamint

[37] https://touche.webis.de/clef24/touche24-web/ideology-and-power-identification-in-parliamentary-debates.html

[38] In particular with "2023 SPEED CHES – Ukraine" and "2020 SPEED CHES – Covid", cf. https://www.chesdata.eu/ches-europe

[39] https://v-dem.net/

Another, easier approach might be to develop "ParlaMint-light" corpora, i.e. corpora that are ParlaMint-encoded, and can therefore take advantage of the validation and conversion software, but might be lacking much of the metadata or annotations. This could be achieved relatively easily by relaxing the validation procedure, developing scripts to convert existing parliamentary corpora to ParlaMint, manually adding only basic information about the parliament, and annotating the transcripts with e.g. UD-Pipe. This light approach could be applied to some of the European countries missing from ParlaMint but for which corpora already exist, such as Germany (Blätte & Blessing, 2018), Ireland (Sylvester et al., 2022) or Slovakia (Mochtak, 2022). Such corpora would not be as richly annotated as the current crop but could nevertheless be a valuable addition to ParlaMint.

Still, probably currently, the most important part of future work does not concern the enhancement of the corpora but encouraging their use, esp. in the disciplines where the use of general purpose corpora is still rare, such as in political science or history.

Acknowledgements. The authors are indebted to the following collaborators, who contributed to the compilation of the ParlaMint corpora: Martin Kirnbauer, Hannes Pirker, Daniel Schopper (AT); Michal Mochtak (BA, HR, RS); Griet Depoorter, Henk van der Pol (BE); Jesse de Does (BE, NL); Ilko Grigorov, Vladislava Grigorova (BG); Dorte Haltrup Hansen, Bart Jongejan (DK); Kadri Vider (EE); Monica Albini, Maria del Mar Bonet Ramos, Ruben de Libano (ES); Iván Antiba-Cartazo, Marilina Pisani, Rodolfo Zevallos (ES-CT); Mario Barcala, Daniel Bardanca, Elisa Fernández Rei, Marcos Garcia, Noelia García Díaz, Pedro García Louzao, María Pérez Lago, Xosé Luís Regueira, Marta Vázquez Abuín, Adrián Vidal Miguéns, Ainhoa Vivel Couso (ES-GA); Rodrigo Agerri, Manex Agirrezabal, Jon Alkorta, Ekain Arrieta, Kike Fernandez (ES-PV); Laura Sinikallio, Minna Tamper (FI); Giancarlo Luxardo (FR); Matthew Coole, Amanda Nwadukwe (GB); Dimitris Gkoumas, Vassilis Papavassiliou, Stelios Piperidis, Prokopis Prokopidis (GR); Réka Dodé, Kinga Jelencsik-Mátyus, Zsófia Varga (HU); Roberto Battistoni, Francesca Frontini, Carlo Marchetti, Simonetta Montemagni, Valeria Quochi, Manuela Ruisi, Giulia Venturi (IT); Katrien Depuydt, Maarten Marx (NL); Bartłomiej Nitoń, Michał Rudolf (PL); Aida Cardoso, Amália Mendes, Rui Pereira (PT); Johan Jarlbrink, Måns Magnusson, Fredrik Mohammadi Norén (SE); Mihael Ojsteršek, Andrej Pančur (SI); Andriana Rii (UA).

Declarations

Funding: This work was funded by:

- CLARIN ERIC, ParlaMint: Towards Comparable Parliamentary Corpora;
- MT corpora: Oracle Cloud Infrastructure credits were donated in kind to the UCREL NLP research group for the semantic annotation task, and we acknowledge the support of Richard Pitts at Oracle Research.
- Speech corpora: ARIS (Slovenian Research and Innovation Agency), J7-4642 "MEZZANINE - Basic Research for the Development of Spoken Language Resources and Speech Technologies for the Slovenian Language";
- AT: Austrian Academy of Sciences;
- BE: Dutch Language Institute;

- BG: Bulgarian Ministry of Education and Science, DO1-301/17.12.21: “Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favour of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH”;
- CZ: Ministry of Education, Youth and Sports of the Czech Republic, LM2023062: "LINDAT/CLARIAH-CZ: Digital Research Infrastructure for Language Technologies, Arts and Humanities";
- DK: Department of Nordic Studies and Linguistics (NorS), University of Copenhagen CLARIN-DK, "CLARIN-DK";
- ES: Spanish Ministry of Science and Innovation, project "Original, translated and interpreted representations of the refugee cris(e)s: methodological triangulation within corpus-based discourse studies" (PID2019-108866RB-I00 / AEI / 10.13039/501100011033)
- ES-GA: Galician Language Institute, University of Santiago de Compostela;
- ES-GA: Xunta de Galicia – University of Santiago de Compostela, 2021-CP080: "Nós: Galician in the society and economy of artificial intelligence (2021-CP080), an agreement between Xunta de Galicia and the University of Santiago de Compostela";
- ES-PV: HiTZ – Ixa Group (UPV/EHU);
- ET: The Postimees Foundation grant "Developments and contradictions on the Estonian political landscape 2011–2023 through the analysis of big textual data".
- FI: Research Council of Finland: DIGIHUM research programme project "Semantic Parliament: Linked Open Data Service for Studying Political Culture (SEMPARL)" (2020–2022), FIN-CLARIAH, part of the EU infrastructures CLARIN and DARIAH;
- GR: Institute for Language and Speech Processing / ATHENA Research Centre
- HU: Hungarian Research Centre for Linguistics;
- IS: The Árni Magnsússon Institute for Icelandic Studies;
- LV: The EU Recovery and Resilience Facility project "Language technology Initiative" (2.3.1.1.i.0/1/22/I/CFLA/002);
- NL: Nederlandse Organisatie voor Wetenschappelijk Onderwijs, CISC.CC.016: "Access to City Councils using Exploratory Search Systems";
- NO: National Library of Norway;
- PL: European Commission programme, POIR.04.02.00-00C002/19: European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure"; co-financed by the Polish Ministry of Education and Science under the agreement 2022/WK/09;
- PL: Institute of Computer Science, Polish Academy of Sciences – "statutory research";
- PL: Polish Ministry of Education and Science, 2022/WK/09: "National contribution to CLARIN ERIC – European Research Infrastructure Consortium:

Common Language Resources and Technology Infrastructure 2022–2023 (CLARIN Q)";

- PT: Fundação para a Ciência e a Tecnologia, UIDP/00214/2020;
- SI, BA, HR, SR, UA: Research infrastructure CLARIN.SI (Jožef Stefan Institute);
- SI: ARIS (Slovenian Research and Innovation Agency), P2-0103: "Knowledge Technologies";
- SI: ARIS (Slovenian Research and Innovation Agency), P6-0411: "Language Resources and Technologies for Slovene";
- SI: ARIS (Slovenian Research and Innovation Agency), P6-0436 "Basic national research program "Digital Humanities";
- SI, ParlaSent: ARIS (Slovenian Research and Innovation Agency), N6-0099 "Flemish-Slovenian bilateral basic research project 'Linguistic landscape of hate speech online' (2019–2023)";
- UA: ARIS (Slovenian Research and Innovation Agency), N6-0288 "The MSCA Seal of Excellence postdoctoral project 'The Changing Discursive Semantics of EU Representations' (2022–2024)";

Conflict of interest/Competing interests: The authors have no conflict of interest, nor competing interest to disclose, neither financial nor any other. One of the authors is a member of the editorial board of this journal.

Ethics approval: Not applicable / The provider of this data and related work declares that, to the best of their knowledge, it is free of copyright restrictions and does not contain sensitive personal information or violate privacy laws.

Consent to participate: No human subjects were involved in this work.

Consent for publication: All authors and other individuals, associated with the work described give their consent to the publication of the article.

Availability of data and materials: The research data described in this paper are available for download under one of the Creative Commons licences.

Code availability: The code and other components associated with the work described in this article are available via the project's Github repository as Open Source.

Authors' contributions:

- T. Erjavec wrote Sections 1, 2 (except for 2.3), and most of 5. In the project, he was the co-leader of the work package WP1: Documentation, interoperability, metadata and lead of WP2: Corpus expansion. He performed most of the work in the tasks T1.1: Harmonisation of encoding and T2.3: Data distribution and oversaw the work in T1.3: Adding metadata to existing corpora, T2.1: Adding new corpora and T2.2: Extending existing corpora.
- M. Kopp prepared the Figures and Tables in the paper. He was the co-leader of the work package WP1: Documentation, interoperability, metadata, where he was involved in tasks T1.1: Harmonisation of encoding. He performed most of the work in task T1.2: Git management and oversaw tasks T2.1: Adding new corpora and T2.2: Extending existing corpora. He produced the ParlaMint-CZ and ParlaMint-UA corpora and contributed to ParlaMint-ES.

- N. Ljubešić wrote Section 4.4. He was the lead of WP3: Corpus enrichment and led task T.3.2: Multimodality. He was involved in the production of the ParlaMint-BA, ParlaMint-BG, ParlaMint-HR, ParlaMint-SI, and ParlaMint-RS corpora.
- T. Kuzman wrote the machine translation part of Section 4.3. She performed the complete MT, as part of T3.1 Machine translation and semantic tagging.
- P. Rayson wrote the semantic annotation part of Section 4.3. He led the semantic annotation task in T3.1 Machine translation and semantic tagging. He led the production of the ParlaMint-GB corpus.
- P. Osenova wrote the dissemination part of Section 5. She was the co-leader of WP5: Coordination and was central to the tasks T5.1: Management and T5.2: Dissemination. She led the production of the ParlaMint-BG corpus.
- M. Ogrodniczuk was the co-leader of WP5: Coordination and was central to the tasks T5.1: Management. He led the production of the ParlaMint-PL corpus.
- Ç. Çöltekin was the co-leader of WP4: Engagement activities and had the sole responsibility for T4.3: Shared task. He led the production of the ParlaMint-TR corpus.
- D. Koržinek performed a significant part of the work on task T3.2: Multimodality.
- K. Meden helped write the Sections 2.2 and 4. In the project, she helped with task T1.1: Harmonisation of encoding and contributed to task T1.3: Adding metadata to existing corpora. She led the production of the ParlaMint-SI corpus.
- J. Skubic performed most of the work in T1.3: Adding metadata to existing corpora.
- P. Rupnik compiled the ParlaMint-BA, ParlaMint-HR, and ParlaMint-RS corpora and participated in the work on task T3.2: Multimodality.
- J. Vidler performed the semantic annotation task in T3.1 Machine translation and semantic tagging.
- The other authors wrote the part of the paper that pertains to their corpus and compiled the individual corpora.
- and D. Fišer was the co-leader of WP4: Engagement activities, and centrally contributed to T4.1: Tutorial and T4.2: Hackathon. She was also the driving force behind the ParlaMint projects.

References

Agnoloni, T., Bartolini, R., Frontini, F., Montemagni, S., Marchetti, C., Quochi, V., … Venturi, G. (2022). Making Italian parliamentary records machine-actionable: the construction of the ParlaMint-IT corpus. *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation conference* (pp. 117–124). Retrieved from https://aclanthology.org/2022.parlaclarin-1.17

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., Vollgraf, R. (2019). FLAIR: An easy-to-use framework for state-of-the-art NLP. *NAACL 2019,* *2019 annual conference of the North American chapter of the Association for Computational Linguistics (demonstrations)* (pp. 54–59).

Alexander, M., & Struan, A. (2022). "In barbarous times and in uncivilized countries" two centuries of the evolving uncivil in the Hansard Corpus. *International Journal of Corpus Linguistics*, *27*(4), 480–505,

Behzad, S., & Zeldes, A. (2020). A cross-genre ensemble approach to robust Reddit part of speech tagging. *Proceedings of the 12th web as corpus workshop (WAC-XII)* (pp. 50–56).

Blätte, A., & Blessing, A. (2018, May). The GermaParl corpus of parliamentary protocols. N. Calzolari et al. (Eds.), *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018).* Miyazaki, Japan: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L18-1130

Branco, A., & Silva, J. (2004, May). Evaluating solutions for the rapid development of state-of-the-art POS taggers for Portuguese. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, & R. Silva (Eds.), *Proceedings of the fourth international conference on language resources and evaluation (LREC'04).* Lisbon, Portugal: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2004/pdf/572.pdf

Branco, A., Silva, J.R., Gomes, L., António Rodrigues, J. (2022, June). Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. N. Calzolari et al. (Eds.), *Proceedings of the thirteenth language resources and evaluation conference* (pp. 5617–5626). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.lrec-1.603

Çöltekin, Ç. (2010). A freely available morphological analyzer for Turkish. *Proceedings of the 7th international conference on language resources and evaluation (LREC 2010)* (pp. 820–827). Retrieved from http://www.lrec-conf.org/proceedings/lrec2010/summaries/109.html

Çöltekin, Ç., Kopp, M., Morkevičius, V., Ljubešić, N., Meden, K., Erjavec, T. (2024, January). *Training data for the shared task Ideology and Power Identification in Parliamentary Debates.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.10450640

Coppedge, M., Gerring, J., Glynn, A., Knutsen, C.H., Lindberg, S.I., Pemstein, D., … et al. (2020). *Varieties of democracy: Measuring two centuries of political change.* Cambridge University Press.

de Marneffe, M.-C., Manning, C.D., Nivre, J., Zeman, D. (2021, 07). Universal Dependencies. *Computational Linguistics*, *47*(2), 255-308, https://doi.org/10.1162/coli_a_00402 Retrieved from https://doi.org/10.1162/coli_a_00402 (https://doi.org/10.1162/coli_a_00402)

Erjavec, T., Kopp, M., Meden, K. (2024). Experience of remote collaborative work in the ParlaMint project using git. *Proceedings of the TwinTalks Workshop at DH2023 (in print).* Graz, Austria: CEUR. (https://ceur-ws.org/)

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agerri, R., Agirrezabal, M., . . . Fišer, D. (2023). *Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 4.0.* Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1860 (http://hdl.handle.net/11356/1860)

Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., Agirrezabal, M., Agnoloni, T., . . . Fišer, D. (2023). *Multilingual comparable corpora of parliamentary debates ParlaMint 4.0.* Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1859 (http://hdl.handle.net/11356/1859)

Erjavec, T., Meden, K., Skubic, J. (2023). Adding political orientation metadata to ParlaMint corpora. *CLARIN annual conference 2023, book of abstracts.* (https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf)

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., . . . Fišer, D. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, *57*(1), 415-448, https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T., & Pančur, A. (2022). The Parla-CLARIN recommendations for encoding corpora of parliamentary proceedings. *Journal of the Text Encoding Initiative (Selected Papers from the 2019 TEI Conference)*(14), , https://doi.org/10.4000/jtei.4133

Fišer, D., & Lenardič, J. (2018). CLARIN corpora for parliamentary discourse research. *Proceedings of the LREC 2018 workshop ParlaCLARIN: Creating and using parliamentary corpora.* European Language Resources Association. (http://lrec-conf.org/workshops/lrec2018/W2/summaries/14_W2.html)

Grünewald, S., Friedrich, A., Kuhn, J. (2021, August). Applying Occam's razor to transformer-based dependency parsing: What works, what doesn't, and what is really necessary. S. Oepen, K. Sagae, R. Tsarfaty, G. Bouma, D. Seddah, & D. Zeman (Eds.), *Proceedings of the 17th international conference on parsing* *technologies and the IWPT 2021 shared task on parsing into enhanced Universal Dependencies (IWPT 2021)* (pp. 131–144). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.iwpt-1.13

Guðjónsson, Á.A., Loftsson, H., Daðason, J.F. (2021). *Icelandic NER API - ensamble model (21.09).* Retrieved from http://hdl.handle.net/20.500.12537/159 (CLARIN-IS)

Hladká, B., Kopp, M., Straňák, P. (2020). Compiling Czech parliamentary stenographic protocols into a corpus. *Proceedings of the LREC 2020 workshop on creating, using and linking of parliamentary corpora with other types of political discourse (ParlaCLARIN II)* (pp. 18–22). Paris, France: European Language Resources Association (ELRA). (https://www.aclweb.org/anthology/2020.parlaclarin-1.4)

Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A. (2020). *spaCy: Industrial-strength natural language processing in Python.*

Janssen, M. (2016, May). TEITOK: Text-faithful annotated corpora. N. Calzolari et al. (Eds.), *Proceedings of the tenth international conference on language resources and evaluation (LREC'16)* (pp. 4037–4043). Portorož, Slovenia: European Language Resources Association (ELRA). Retrieved from https://aclanthology.org/L16-1637

Jasonarson, A., Steingrímsson, S., Sigurðsson, E.F., Daðason, J.F. (2022). *COMBO-based UD parser 22.10.* Retrieved from http://hdl.handle.net/20.500.12537/272 (CLARIN-IS)

Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., … Vachudova, M.A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies*, *75*, 102420, https://doi.org/https://doi.org/10.1016/j.electstud.2021.102420 Retrieved from https://www.sciencedirect.com/science/article/pii/S0261379421001323

Jongejan, B., & Dalianis, H. (2009, August). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. K.-Y. Su, J. Su, J. Wiebe, & H. Li (Eds.), *Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP* (pp. 145–153). Suntec, Singapore: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P09-1017

Junczys-Dowmunt, M., Grundkiewicz, R., Dwojak, T., Hoang, H., Heafield, K., Neckermann, T., … Birch, A. (2018, 07). Marian: Fast neural machine translation in C++. *Proceedings of ACL 2018, system demonstrations* (pp. 116–121). Melbourne, Australia: Association for Computational Linguistics. Retrieved from http://www.aclweb.org/anthology/P18-4020

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., ... Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*, 7–36,

Kondratyuk, D., & Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 2779–2795). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1279

Kopp, M. (2022). *ParCzech pipeline.* https://github.com/ufal/ParCzech.

Kopp, M. (2024a). *AudioPSP 24.01: Audio recordings of proceedings of the chamber of deputies of the parliament of the Czech Republic.* LINDAT/CLARIAH-CZ digital library. (http://hdl.handle.net/11234/1-5404)

Kopp, M. (2024b). *ParCzech 4.0.* LINDAT/CLARIAH-CZ digital library. (http://hdl.handle.net/11234/1-5360)

Kopp, M., Stankov, V., Krůza, J., Straňák, P., Bojar, O. (2021). ParCzech 3.0: A large Czech speech corpus with rich metadata. K. Ekštein, F. Pártl, & M. Konopík (Eds.), *Text, speech, and dialogue* (pp. 293–304). Cham, Switzerland: Springer. (https://doi.org/10.1007/978-3-030-83527-9_25)

Koržinek, D., & Ljubešić, N. (2024). *Parliamentary spoken corpus of Polish ParlaSpeech-PL 1.0.* Slovenian language resource repository CLARIN.SI. (http://hdl.handle.net/11356/1686)

Kryvenko, A., Evkoski, B., Bordon, D., Meden, K. (2023). Splitting lips: polarization through parliamentary speech. *Poster presented at the Helsinki digital humanities hackathon #DHH23.* (https://www.helsinki.fi/assets/drupal/2023-06/dhh23-parliament-poster.pdf)

Kryvenko, A., & Kopp, M. (2023). Workflow and metadata challenges in the ParlaMint project: Insights from building the ParlaMint-UA corpus. *CLARIN annual conference proceedings 2023* (pp. 67–70). Leuven, Belgium: CLARIN ERIC. (https://office.clarin.eu/v/CE-2023-2328_CLARIN2023_ConferenceProceedings.pdf)

Kryvenko, A., & Pahor de Maiti, K. (2023). *Combining corpus linguistics and discourse analysis to explore the parliamentary debates across Europe.* Retrieved from https://digihubb.centre.ubbcluj.ro/workshops/ (Tutorial given at the European Summer University in Digital Humanities, Babeș-Bolyai University, Cluj-Napoca, Romania)

Kryvenko, A., Pahor de Maiti, K., Osenova, P. (2023). *Put Them In to Get Them Out: the ParlaMint Corpora for Digital Humanities and Social Sciences Research.* Retrieved from https://dh2023.adho.org/?page_id=616 (Tutorial given at the Digital Humanities conference 2023, Graz)

Kuzman, T., Ljubešić, N., Erjavec, T., Kopp, M., Ogrodniczuk, M., Osenova, P., . . . Fišer, D. (2023). *Linguistically annotated multilingual comparable corpora of parliamentary debates in English ParlaMint-en.ana 4.0.* Slovenian language resource repository CLARIN.SI. (http://hdl.handle.net/11356/1864)

Laur, S., Orasmaa, S., Särg, D., Tammo, P. (2020, May). EstNLTK 1.6: Remastered Estonian NLP pipeline. *Proceedings of the 12th language resources and evaluation conference* (pp. 7154–7162). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.884

Lenardič, J., & Fišer, D. (2023). *CLARIN resource families: Parliamentary corpora.* (https://www.clarin.eu/resource-families/parliamentary-corpora, accessed on 2024-01-20.)

Ljubešić, N., & Dobrovoljc, K. (2019, August). What does neural bring? analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. *Proceedings of the 7th workshop on Balto-Slavic natural language processing* (pp. 29–34). Florence, Italy: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/W19-3704

Ljubešić, N., Koržinek, D., Rupnik, P. (2024). *Parliamentary spoken corpus of Croatian ParlaSpeech-HR 2.0.* Slovenian language resource repository CLARIN.SI. (http://hdl.handle.net/11356/1914)

Ljubešić, N., Koržinek, D., Rupnik, P., Jazbec, I.-P. (2022, June). ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. D. Fišer, M. Eskevich, J. Lenardič, & F. de Jong (Eds.), *Proceedings of the workshop ParlaCLARIN III within the 13th language resources and evaluation conference* (pp. 111–116). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2022.parlaclarin-1.16

Ljubešić, N., Koržinek, D., Rupnik, P., Jazbec, I.-P., Batanović, V., Bajčetić, L., Evkoski, B. (2022). *ASR training dataset for Croatian ParlaSpeech-HR v1.0.* Retrieved from http://hdl.handle.net/11356/1494 (Slovenian language resource repository CLARIN.SI)

Ljubešić, N., Rupnik, P., Koržinek, D. (2024). *Parliamentary spoken corpus of serbian ParlaSpeech-RS 1.0.* Slovenian language resource repository CLARIN.SI. (http://hdl.handle.net/11356/1834)

Machálek, T. (2020, May). KonText: Advanced and flexible corpus query interface. *Proceedings of the 12th language resources and evaluation conference* (pp. 7003–7008). Marseille, France: European Language Resources Association. Retrieved from https://www.aclweb.org/anthology/2020.lrec-1.865

Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. *Association for Computational Linguistics (ACL) system demonstrations* (pp. 55–60). Retrieved from http://www.aclweb.org/anthology/P/P14/P14-5010

Mochtak, M. (2022, August). *SVKCorp: Corpus of debates in the national council of the Slovak Republic.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.7020534

Monarch, R., & Munro, R. (2021). *Human-in-the-loop machine learning: Active learning and annotation for human-centered AI.* Simon and Schuster.

Nivre, J., Agić, Ž., Ahrenberg, L., Aranzabe, M.J., Asahara, M., Atutxa, A., ... Zhu, H. (2017). *Universal Dependencies 2.0.* Retrieved from http://hdl.handle.net/11234/1-1983 (LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University)

Orosz, G., Szántó, Z., Berkecz, P., Szabó, G., Farkas, R. (2022). HuSpaCy: an industrial-strength Hungarian natural language processing toolkit. *XVIII. Magyar Számítógépes Nyelvészeti Konferencia* (pp. 59–73).

Pančur, A., Erjavec, T., Meden, K., Ojsteršek, M., Šorn, M., Blaj Hribar, N. (2022). *Slovenian parliamentary corpus (1990-2022) siParl 3.0.* Slovenian language resource repository CLARIN.SI. Retrieved from http://hdl.handle.net/11356/1748 (http://hdl.handle.net/11356/1748)

Pančur, A., & Erjavec, T. (2020, May). The siParl corpus of Slovene parliamentary proceedings. D. Fišer, M. Eskevich, & F. de Jong (Eds.), *Proceedings of the second ParlaCLARIN workshop* (pp. 28–34). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.parlaclarin-1.6

Prokopidis, P., & Piperidis, S. (2020). A neural NLP toolkit for Greek. *11th Hellenic conference on artificial intelligence* (pp. 125–128).

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D. (2020). Stanza: A Python natural language processing toolkit for many human languages. *Proceedings of the 58th annual meeting of the association for computational linguistics: System demonstrations.*

Rayson, P., Archer, D., Piao, S., McEnery, T. (2004). The ucrel semantic analysis system. *Proceedings of the workshop on beyond named entity recognition semantic labelling for nlp tasks, in association with lrec-04* (pp. 7–12).

Silveira, N., Dozat, T., de Marneffe, M.-C., Bowman, S., Connor, M., Bauer, J., Manning, C.D. (2014). A gold standard dependency corpus for English. *Proceedings of the ninth international conference on language resources and evaluation (LREC-2014).*

Skubic, J., Angermeier, J., Bruncrona, A., Evkoski, B., Leiminger, L. (2022). Networks of power: Gender analysis in selected European parliaments. *Proceedings of the 2nd workshop on computational linguistics for political text analysis (CPSS-2022).* (https://old.gscl.org/en/arbeitskreise/cpss/cpss-2022/workshop-proceedings-2022)

Steingrímsson, S., Barkarson, S., Örnólfsson, G.T. (2020, May). IGC-Parl: Icelandic corpus of parliamentary proceedings. D. Fišer, M. Eskevich, & F. de Jong (Eds.), *Proceedings of the second ParlaCLARIN workshop* (pp. 11–17). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.parlaclarin-1.3

Stopfner, M. (2018). Put your "big girl" voice on: Parliamentary heckling against female MPs. *Journal of Language and Politics*, *17*(5), 617–635,

Straka, M. (2018, October). UDPipe 2.0 prototype at CoNLL 2018 UD shared task. *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 197–207). Brussels, Belgium: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/K18-2020

Straková, J., Straka, M., Hajič, J. (2019). Neural architectures for nested NER through linearization. *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 5326–5331). Stroudsburg, PA, USA: Association for Computational Linguistics.

Sylvester, C., Greene, Z., Ebing, B. (2022). *ParlEE plenary speeches data set: Annotated full-text of 21.6 million sentence-level plenary speeches of eight EU states.* (https://doi.org/10.7910/DVN/ZY3RV7, Harvard Dataverse, V1)

Tamper, M., Leskinen, P., Apajalahti, K., Hyvönen, E. (2018, November). Using biographical texts as linked data for prosopographical research and applications. M. Ioannides et al. (Eds.), *Digital heritage. progress in cultural heritage: Documentation, preservation, and protection. 7th international conference, EuroMed 2018* (pp. 125–137). Nicosia, Cyprus: Springer-Verlag.

Tamper, M., Oksanen, A., Tuominen, J., Hietanen, A., Hyvönen, E. (2020). Automatic annotation service APPI: Named entity linking in legal domain. A. Harth et al. (Eds.), *The semantic web: ESWC 2020 satellite events* (Vol. 12124, pp. 208–213). Springer-Verlag. Retrieved from https://doi.org/10.1007/978-3-030-62327-2_36

TEI Consortium (Ed.). (2017). *TEI P5: Guidelines for electronic text encoding and interchange.* TEI Consortium. Retrieved from http://www.tei-c.org/Guidelines/P5/

Terčon, L., & Ljubešić, N. (2023). *CLASSLA-Stanza: The next step for linguistic processing of South Slavic languages.*

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. *LREC'12* (Vol. 2012, pp. 2214–2218).

Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – building open translation services for the world. *Proceedings of the 22nd annual conference of the European Association for Machine Translation.*

Tjong Kim Sang, E.F., & De Meulder, F. (2003). Introduction to the CoNLL-2003 shared task: language-independent named entity recognition. *Proceedings of the seventh conference on natural language learning at HLT-NAACL 2003 - volume 4* (p. 142–147). USA: Association for Computational Linguistics. Retrieved from https://doi.org/10.3115/1119176.1119195

Truan, N. (2019). Talking about, for, and to the people: Populism and representation in parliamentary debates on Europe. *Zeitschrift für anglistik und amerikanistik*, *67*(3), 307–337,

Truan, N., & Romary, L. (2022). Building, encoding, and annotating a corpus of parliamentary debates in TEI XML: A cross-linguistic account. *Journal of the Text Encoding Initiative*(14), ,

Wissik, T. (2022). Encoding interruptions in parliamentary data: From applause to interjections and laughter. *Journal of the Text Encoding Initiative*(14), , Retrieved from https://doi.org/10.4000/jtei.4214

Wissik, T., & Pirker, H. (2018). ParlAT beta corpus of Austrian parliamentary records. *Proceedings of the LREC 2018 workshop ParlaCLARIN: Creating and using parliamentary corpora.*

Zeldes, A. (2017). The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, *51*(3), 581–612, https://doi.org/http://

dx.doi.org/10.1007/s10579-016-9343-x

Znotins, A., & Cirule, E. (2018). NLP-PIPE: Latvian NLP tool pipeline. *Human language technologies - the Baltic perspective* (Vol. 307, p. 183-189). IOS Press. Retrieved from http://ebooks.iospress.nl/volumearticle/50320