Proceedings of ParlaCLARIN III @ LREC2022 pages 1–6
Marseille, 20 June 2022

ParlaMint II: The Show Must Go On

Maciej Ogrodniczuk, Petya Osenova, Tomaž Erjavec, Darja Fišer, Nikola Ljubešić, Çagrı Çöltekin, Matyáš Kopp, Katja Meden

maciej.ogrodniczuk@ipipan.waw.pl,
petya@bultreebank.org, tomaz.erjavec@ijs.si, darja.fiser@ff.uni-lj.si,
nikola.ljubesic@ijs.si, ccoltekin@sfs.uni-tuebingen.de,
kopp@ufal.mff.cuni.cz, katja.meden@ijs.si

Abstract

In ParlaMint I, a CLARIN-ERIC supported project, a set of comparable and uniformly annotated multilingual corpora for 17 national parliaments was developed and released. Currently on-going is the ParlaMint II project, where the main goals are to upgrade the annotation guidelines, XML schema and Git-related workflow; enhance the existing corpora with new metadata and newer data; add corpora for 10 new parliaments; add machine-translated and semantically annotated English texts to the corpora; for a few corpora add speech data; and provide more use cases. The paper reports on these planned steps, including some that have already been taken, and outlines future plans.

Keywords: parliamentary debates, parliamentary records, parliamentary corpora, ParlaMint, linguistic annotation, metadata

1. Introduction

The ParlaMint project produced uniformly sampled, annotated and encoded comparable parliamentary corpora for 17 European countries with almost half a billion words in total (Erjavec et al., 2022). The corpora, which comprise reference and COVID-19 sections, contain rich metadata about the mandates, sessions, and speakers and their political party affiliations etc., are linguistically annotated for named entities and Universal Dependencies morphological features and syntax, and encoded to a common and very strict schema, so their format is not merely interchangeable but also interoperable. This has been validated in practice, as the corpora have been mounted on the CLARIN.SI concordancers, i.e. they can be explored and analyzed in a common and very powerful environment. The corpus development and a part of the communication took place on GitHub, the corpora have been released under the CC BY licence in the scope of a CLARIN repository (Erjavec et al., 2021a; Erjavec et al., 2021b)[1], not only in their source XML TEI format, but also in a number of derived and immediately useful formats.

The ParlaMint corpora have also been used in the Helsinki Digital Humanities Hackathon (Calabretta et al., 2021)[2], giving them increased visibility as well as providing useful feedback for the structure of the final version 2.1 corpora of the project. The project has thus produced a novel and highly valuable resource for a broad range of comparative trans-national SSH studies that is openly available and has already proved itself in practice.

However, during the compilation and especially DHH use of the ParlaMint corpora, a number of relatively straightforward as well as some more complex upgrades were identified, which would make the corpora even more useful. In particular, the structure, accessibility and metadata of the corpora need to be improved in order to maximize interoperability and comparative research. These issues are discussed in Section 2.

Due to the prolonged pandemics, the COVID-19 section of the existing corpora also needs to be extended with new data. To make the resource as valuable for SSH scholars as possible, parliamentary corpora of additional countries and languages need to be provided as well. The data extension is the focus of Section 3.

The ParlaMint corpus family will be enriched by adding machine translations into English, thus allowing for comparative analyses across parliaments. Also, integration of speech data will be piloted for selected parliaments. These topics are presented in Section 4.

Last but not least, the corpora will be utilised in new and more varied user scenarios. The engagement activities like the hackathon with the ParlaMint data, as well as the shared task and other related showcases, are outlined in Section 5.

Section 6 concludes and lists some directions to follow beyond the time and resource limits of the project.

2. Schema and Metadata Improvements

The encoding of ParlaMint I corpora followed the previously developed TEI-based Parla-CLARIN recommendations for encoding parliamentary corpora (Erjavec and Pančur, 2019)[3], which provide extensive textual guidelines but are very permissive in their formal

[1] http://hdl.handle.net/11356/1432 and http://hdl.handle.net/11356/1431

[2] https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid-times/

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

XML schema. To enable interoperability of the produced corpora, ParlaMint required much stricter encoding, so we started the project by defining a RelaxNG schema for corpus validation, which was then refined during most of the lifetime of the ParlaMint I project.

However, the Parla-CLARIN textual recommendations were not updated during this time, so Parla-CLARIN was lagging behind ParlaMint. On the other hand, there were no ParlaMint-specific encoding guidelines, and the project partners – those with sufficient digital skills – had to rely on inspecting the formal schema or the already submitted and validated samples of their corpora and try to adapt them to their circumstances. Towards the end of the project, when all corpora were already available, some of the encoding decisions were also discovered to be questionable, however, it was too late to re-encode the corpora by then. Finally, the ParlaMint project had to absorb many corpora in a relatively short time, so many aspects of the encoding were not unified nor harmonised, in particular the status and roles of speakers, the encoding of sessions, meetings, and agendas, the distinction between political parties vs. parliamentary groups etc.

For interoperability of ParlaMint corpora, as well as a step towards standardisation of information (taxonomies/ontologies) associated with parliamentary debates in ParlaMint II, it is was deemed highly beneficial if the encoding and metadata of the current ParlaMint corpora were harmonised and unified, and the Parla-CLARIN recommendations updated to reflect the experience gained in ParlaMint I. This is also strategically important as several corpora being developed independently of the ParlaMint project have already started using the Parla-CLARIN schema, which was the ultimate goal for future sustainability and interoperability of the Parliamentary Resource Family.

ParlaMint II involves more than 30 partners that are all required to submit large and heavily annotated corpora with rich metadata. Yet the corpora will come from completely different sources and embodying different parliamentary procedures and traditions, while the partners will be using different tools for their annotation, and have very different backgrounds and familiarity with TEI, XML, its schema languages, and XSLT. It is, therefore, crucial to establish validation procedures that will result in useful, i.e. correctly and consistently encoded ParlaMint II corpora. Parla-CLARIN, as well as ParlaMint I, have already shown that git is a highly versatile environment for keeping track not only of program code but also documentation. Hosting platforms, in particular GitHub, also provide the means of recorded and structured communication via issues; however, this communication and data exchange policy was not really enforced in ParlaMint I.

Finally, a large part of the value of the ParlaMint corpora comes from their extensive metadata on speakers, which, however, can be very labour intensive to find and add to the corpora for some parliaments, which is why some corpora are missing some metadata that others have. Already in the first use cases, it also turned out that even more metadata would be highly beneficial to enable increasingly wider analyses, as this is unavailable in most related resources.

2.1. Harmonisation of Encoding

In the first phase of the ParlaMint II project, the Parla-CLARIN recommendations were updated adopting the solutions and examples from the ParlaMint corpora, yet still allowing for project-specific extensions.

On the basis of the updated Parla-CLARIN recommendations, but highly specific to ParlaMint, we made a new set of recommendations, including the schema (i.e. a XML TEI ODD document), and made the text guidelines available via GitHub pages[4]. These guidelines are meant to serve as the basis for adding new corpora and extending the existing ones. It should be noted that while the ParlaMint RelaxNG schemas derived from the ParlaMint ODD can be used for validation, more precise validation is still achieved with the native ParlaMint RelaxNG schemas and even more with developed validating XSLT scripts.

The encoding of ParlaMint corpora was further unified as regards the use of attributes and their values, including legislative taxonomies and vocabularies (mostly in both source language and English), speaker roles and affiliations etc. The existing ParlaMint schema and corpora were modified to reflect the ParlaMint best practice. In the course of these modifications, all observed open problems were documented via project GitHub issues.

2.2. Git Management

Both Parla-CLARIN and ParlaMint are completely or largely hosted on GitHub, and both need to be updated in a harmonised and controlled fashion, also providing support for the existing and new corpora developers, as well as to the use cases working with the data. In ParlaMint II, we have already improved the usage of GitHub, e.g. we implemented much stricter validation of commits, encoding and statistical documentation of the corpora in HTML, regular milestones and releases, and are responsive to communication via issues. While in ParlaMint I, quite a lot of support was done via individual emails, this is, given the even larger number of partners in ParlaMint II, now unmanageable, so all communication is to be via GitHub issues, and the validation of corpora the direct responsibility of the partners.

2.3. Adding Metadata to Existing Corpora

Additional metadata, in particular the information about whether the speakers are members of the government (ministers), and the positioning of political parties on the left-right spectrum, are planned to be added

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

to the corpora in ParlaMint II. The partners who will encode this information in their corpora can take advantage of a pipeline that transforms the data entered in spreadsheets into the required XML encoding. In addition to this, a taxonomy of common ministry types is planned to be added to enable cross-corpus comparisons.

After the initial discussion, it became clear that these tasks might present us with some difficulties, stemming mainly from the fact that the structure of political systems differs severely from one country to another. Primary tasks for this section are therefore finding common ground to facilitate encoding of the additional information about members of the government (and the taxonomy) as well as finding the appropriate scale for encoding political orientation (left-right scale, political compass scale or other).

3. Corpus Expansion

New corpora will be added to ParlaMint, and existing corpora will be updated with newer materials. All new resources will contain material from the same minimum periods, the same metadata as existing ParlaMint corpora and linguistic annotations. This will extend the ParlaMint scope in countries, languages and time to make it even more interesting for researchers.

As with previous versions of ParlaMint, both new and updated corpora will be validated, converted to derived formats (plain text, metadata files, CoNLL-U, vertical files), mounted on the CLARIN.SI concordancers, and deposited in the CLARIN.SI repository. We envision three releases: 3.0 at the half-way mark, 3.1 shortly before the end, and 3.2 at the conclusion of the project.

3.1. Adding New Corpora

New parliamentary corpora will be prepared by 10 new project partners (from Austria, Basque Country, Catalonia, Estonia, Finland, Greece, Norway, Portugal, Romania and Sweden) according to ParlaMint specifications and guidelines. The parliamentary transcripts will cover the period at least between January 1, 2015, and February 1, 2022. The texts in the corpora will be split into reference (until October 31, 2019) and COVID parts (later data).

Following the ParlaMint I model, each corpus will have to be delivered in two variants, the TEI encoded plain text one with the metadata and transcripts of the speeches, and the linguistically annotated one (so-called TEI.ana) with added linguistic annotations.

Corpus metadata should contain at least type of parliament (unicameral, bicameral), which speeches are included (lower/upper house, mandates) and the structure of the proceedings (taxonomy with types of meetings, types of speakers, legislative periods). Corpus element structure should encompass date-stamped mandates, sessions and speeches. Each speech can, minimally, contain only the pure transcripts of the speeches divided into paragraphs. However, many transcripts also contain commentary by the transcribers, which are then also retained and encoded.

Metadata on speakers should contain speaker role (regular, chair, guest), their analysed name (forename, surname), gender, MP status and political affiliation(s). If their MP status and political affiliation changed in the time frame of the corpus, it needs to be time-stamped. Political parties and/or political groups should be marked with name, short name (initials) and possibly start/end of existence. Coalitions/oppositions of parties should also be marked in the time frame of the corpus.

A linguistic annotation should encompass tokenisation and sentence segmentation, lemmatisation and UD morphological features, UD syntactic annotations and NE marking (PER, LOC, ORG, MISC).

3.2. Extending Existing Corpora

Recent data (up to June 2022) will also be added to the existing ParlaMint dataset, and the corpora will be further fixed for the found errors and missing metadata. Concerning COVID, it is difficult at this point to consider what period would be viewed as in-pandemic and post-pandemic but we can update our categorization accordingly later.

4. Corpus Enrichment

In this task, we will enhance the ParlaMint corpora with a translation of all non-English transcriptions into English. Having all the corpora in English will enable treating them as one corpus, and using identical queries to view and analyse the data from various parliaments. This opens the way for simple translingual comparative analyses among more than 20 national and regional parliaments, importantly increasing the usability of the ParlaMint corpora, making them an even more globally relevant research dataset. Furthermore, we will semantically tag the translated corpus, which will significantly increase the value of the ParlaMint corpora for SSH scholars.

As a proof of concept, we will also add a subset of recordings of parliamentary debates for selected parliaments, and align them with the available transcriptions. This will have a multiplier effect on improving interoperability of speech / multimodal data and tools in CLARIN well beyond parliamentary records, and will enable novel dimensions of SSH research on ParlaMint corpora that is currently not possible with most related resources, such as comparisons of official transcriptions with actual parliamentary discussions within and across parliaments.

4.1. Machine Translation

As part of the preparation of the DHH 2021 hackathon, we already machine translated 10 of the ParlaMint I corpora to English using the OpenNMT system (Klein et al., 2017)[5]. Now we plan to machine translate all

[5] https://opennmt.net/

the ParlaMint II corpora to English, with the best-performing model at the time of the translation task. We also plan to explore automatic post-editing to fix the most frequent translation errors, in particular the frequently incorrect "translation" of names. The translations will be performed on the sentence level, so that the sentence-level alignment between the originals and translations to English are available.

4.2. Semantic Tagging

The resulting texts will then be encoded and linguistically annotated in the same way as the other corpora. We will also add semantic annotations to the translated corpora using the UCREL Semantic Analysis System, USAS (Rayson et al., 2004)[6], which has been developed by the U.K. ParlaMint partner. The USAS tagger will assign semantic fields from a taxonomy of 232 tags to words and multi-word expressions in the corpus, representing coarse-grained word senses. The system for English has been developed and applied over the last 30 years, and assigning semantic tags to the English translations (as well as the original UK subcorpus) will facilitate future work on bootstrapping prototype semantic taggers in other languages via sentence and word alignment. USAS tagging accuracy for English is 91% and the tagger is already freely available via the UCREL website and REST API. In parallel, the Lancaster UCREL centre will develop a Python open-source version of the USAS tagger.

4.3. Multimodality

We will also gather, process and align audio recordings with the transcriptions for a selected list of languages. The alignments will be performed on the level of segments lasting 5 to 30 seconds. The possibility of making the aligned audio available through the KonText concordancer (Machálek, 2020) will be investigated as well.

Due to the high technical complexity of this task, it will be run as a proof-of-concept on three selected languages (Czech, Polish and Croatian), where audio alignment activities have already been applied to some level. Each of the selected languages will deliver at least 50 hours of high-quality audio alignment as well as the code base and a report of the used alignment procedure. The aim of this task is not only to obtain aligned audio data for the selected the ParlaMint corpora, but to identify best practices in the currently highly vibrant area of speech processing, to be used on the remaining ParlaMint languages in a possible follow-up project.

Although the project has started only recently, we can already report on the first freely-available dataset for training automatic-speech-recognition systems for Croatian, ParlaSpeech-HR (Ljubešić et al., 2022). It is based on the ParlaMint I corpus and the available video recordings of the Croatian parliament, resulting in a dataset of 1,816 hours. A similar availability of the speech and transcript data will be ensured for the remaining languages as well. The bootstrapping approach to building the dataset, consisting of using Google speech-to-text for constructing an initial dataset, and then training a transformer-based ASR system from this initial dataset, and using it to to build the final dataset, is described in (Ljubešić et al., 2022).

In implementing the alignment of Czech audio and transcription, we plan to utilize tools used to create the ParCzech 3.0 corpus (Kopp et al., 2021). This corpus covers the same period as the ParlaMint I Czech corpus and contains more than 3,000 hours of aligned audio. The audio recordings provided on the Czech Chamber of Deputies web pages are about 14 minutes long with only the approximately middle 10 minutes corresponding to the transcript on one web page, making it difficult to determine the alignment of the audio with the transcript. The alignment algorithm currently used in ParCzech 3.0 does try to determine the beginning of transcription in the audio file but because the transcription is redacted (i.e. does not fully correspond to the audio), the algorithm does not always work correctly. We believe that there is room for improvement by modifying the algorithm to also take into account the alignments made on the previous web page, and we will investigate this upgrade in ParlaMint II.

5. Engagement Activities

5.1. Tutorial

After performing a literature review and interacting with the relevant national and European projects and networks is being documented and made available (Skubic and Fišer, 2022), a tutorial and showcases will be developed for SSH scholars and students which demonstrates the use of ParlaMint data, metadata and linguistic annotations.

The tutorial will be developed around relevant SSH research questions on the theme of "opposition in times of crisis" using topic modelling, one of the most popular methods in the DH community. The tutorial will be complementary to the previous one, *Voices of the Parliament* [7], developed by the same team outside the ParlaMint project, which demonstrates the potential of parliamentary corpora research via concordancers. The new tutorial will be aimed at students and scholars of digital humanities and social sciences who are interested in the study of socio-cultural phenomena through language and to engage with the user-friendly text-mining tool *Orange* [8].

The theoretical part of the tutorial will introduce the characteristics of parliamentary records, the construction of the ParlaMint corpora and topic modelling. The practical part will demonstrate how topic modelling

[6] https://ucrel.lancs.ac.uk/usas/

[7] https://sidih.github.io/voices/index.html

[8] https://orangedatamining.com

and the Orange text mining tool can be utilized to answer three concrete research questions. In Task 1, we will analyze the basic characteristics of parliamentary speeches before and during the pandemic using interactive visualizations. In Task 2, we will identify the central topics of discussions in the two periods. In Task 3, we will explore topic distributions using heatmaps.

5.2. Showcases

Informed by the literature review and interactions with the relevant national and European projects and networks, a collection of showcases will be developed that will demonstrate the value of the ParlaMint corpora for SSH researchers and will serve as an instrument for cross-disciplinary method and knowledge transfer.

5.3. Hackathon

After the successful participation of the ParlaMint community in the Helsinki Digital Humanities Hackathon 2021, ParlaMint corpora will also be used at the DHH Hackathon 2022[9]. This time the participants will focus on the comparison of parliamentary debates from a sociological, politological, and computational perspective. Political decision-making is organised in party groups, committees, and informal networks among members of parliament and civil servants. In the plenary session, we see these networks manifest themselves as speakers represent their respective groups and refer to one another. The degree to which these networks display exceptional polarisation, centralization of parliamentary voices, or an imbalance in the dynamic between government and opposition, is telling of how the principle of parliamentarism is concretely playing out in the different countries. The networks can also be studied from the perspective of gender, party affiliation, and party stability. By comparing the data synchronically and diachronically in a cross-lingual context, we can obtain important insights into transnational characteristics.

This is why the objective of the hackathon will be to learn how to use comparable parliamentary corpora from various European countries that are annotated with rich metadata and linguistic annotations, enabling various analytical directions. The group will take a network analysis perspective on parliament debates to answer questions on the influence of members, the polarisation of groups, and information spreading in parliament. The group will make use of the linguistic annotations, Named Entities, and metadata coded in the ParlaMint data. Additionally, the group will learn to utilise *Google Colab* [10] and network analysis tools such as *Gephi* [11] and *NetworkX* [12] to bring together the disciplines of computer science and humanities in gaining knowledge on the Networks of Power. The results from the hackathon will be published on the CLARIN website.

5.4. Shared Task

To address a very different but important community of users and expose the created resources to novel approaches, a shared task will be organized in which the ParlaMint corpora will be used to predict whether a speech belongs to a governing or opposition party member (and possibly additional tasks for party affiliation and political ideologies). The corpora released in ParlaMint I will be used as training data, and the newly developed but withheld ParlaMint II corpora will be used as test data. The results from the shared task will be published in open-access proceedings. The details about the shared task will be communicated when this information is ready.

6. Beyond ParlaMint II

Even though ParlaMint II will run until 2023, we are already planning how it could be extended in the future and become a sustainable initiative. Apart from such obvious directions as including more data (from the European Parliament, regional parliaments or national parliaments beyond Europe) or historical data, we are also planning to link our datasets with additional data sources, e.g. by adding voting results, referencing social media content or introducing newspaper and TV news mentions.

ParlaMint data could also be extended to include multimodal aligned corpora (with speech and video), gesture annotated corpora or live corpora produced and used as streamed and on the fly.

Acknowledgements

The ParlaMint project is supported by: CLARIN ERIC – 'ParlaMint: Towards Comparable Parliamentary Corpora' • H2020-INFRAEOSC-04-2018 #823782 'SSHOC: Social Sciences and Humanities Open Cloud' • ARRS (Slovenian Research Agency) P2-103 'Knowledge Technologies' • ARRS (Slovenian Research Agency) P6-0411 'Language Resources and Technologies for Slovene' • ARRS (Slovenian Research Agency) P6-0436 'Digital Humanities: resources, tools and methods' • Ministry of Education and Science Republic of Bulgaria DO01-272/16.12.2019 'Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies CLaDA-BG' • LINDAT/CLARIAH-CZ LM2018101 'Digital Research Infrastructure for Language Technologies, Arts and Humanities' • Spanish Ministry of Science and Innovation PID2019-108866RB-I0 / AEI / 10.13039/501100011033 'Original, Translated and Interpreted Representations of the Refugee Cris(e)s: Methodological Triangulation

[9] https://www2.helsinki.fi/en/helsinki-centre-for-digital-humanities/helsinki-digital-humanities-hackathon-2022-dhh22

[10] https://colab.research.google.com

[11] https://gephi.org

[12] https://networkx.org

within Corpus-Based Discourse Studies' • The Research Council of Lithuania P-MIP-20-373 "Policy Agenda of the Lithuanian Seimas and its Framing: The Analysis of the Seimas Debates in 1990 2020" • CLARIN-LV, European Regional Development Fund project 1.1.1.5/18/I/016 'University of Latvia and Institutes in the European Research Area – Excellency, Activity, Mobility, Capacity' • CLARIN-PL-Biz, financed by the European Regional Development Fund as a part of the 2014-2020 Smart Growth Operational Programme, CLARIN – Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00-00C002/19

Bibliographical References

Calabretta, I., Dalton, C., Griscom, R., Kołczyńska, M., Pahor de Maiti, K., and Ros, R. (2021). Parliamentary debates in the COVID times. https://dhhackathon.wordpress.com/2021/05/28/parliamentary-debates-in-the-covid.

Erjavec, T. and Pančur, A. (2019). Parla-CLARIN: TEI guidelines for corpora of parliamentary proceedings, September. https://doi.org/10.5281/zenodo.3446164.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L. D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Darģis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Bartolini, R., Cimino, A., Diwersy, S., Luxardo, G., and Rayson, P. (2021a). Linguistically annotated multilingual comparable corpora of parliamentary debates ParlaMint.ana 2.1. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1431.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Grigorova, V., Rudolf, M., Pančur, A., Kopp, M., Barkarson, S., Steingrímsson, S., van der Pol, H., Depoorter, G., de Does, J., Jongejan, B., Haltrup Hansen, D., Navarretta, C., Calzada Pérez, M., de Macedo, L. D., van Heusden, R., Marx, M., Çöltekin, Ç., Coole, M., Agnoloni, T., Frontini, F., Montemagni, S., Quochi, V., Venturi, G., Ruisi, M., Marchetti, C., Battistoni, R., Sebők, M., Ring, O., Darģis, R., Utka, A., Petkevičius, M., Briedienė, M., Krilavičius, T., Morkevičius, V., Diwersy, S., Luxardo, G., and Rayson, P. (2021b). Multilingual comparable corpora of parliamentary debates ParlaMint 2.1. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1432.

Erjavec, T., Ogrodniczuk, M., Osenova, P., Ljubešić, N., Simov, K., Pančur, A., Rudolf, M., Kopp, M., Barkarson, S., Steingrímsson, S., Çöltekin, Ç., de Does, J., Depuydt, K., Agnoloni, T., Venturi, G., Pérez, M. C., de Macedo, L. D., Navarretta, C., Luxardo, G., Coole, M., Rayson, P., Morkevičius, V., Krilavičius, T., Darģis, R., Ring, O., van Heusden, R., Marx, M., and Fišer, D. (2022). The ParlaMint Corpora of Parliamentary Proceedings. *Language Resources and Evaluation*. https://doi.org/10.1007/s10579-021-09574-0.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada, July. Association for Computational Linguistics. https://www.aclweb.org/anthology/P17-4012.

Kopp, M., Stankov, V., Krůza, J. O., Straňák, P., and Bojar, O. (2021). ParCzech 3.0: A Large Czech Speech Corpus with Rich Metadata. In *24th International Conference on Text, Speech and Dialogue*, pages 293–304, Cham, Switzerland. Springer. https://link.springer.com/chapter/10.1007/978-3-030-83527-9_25.

Ljubešić, N., Korzinek, D., Rupnik, P., and Jazbec, I.-P. (2022). ParlaSpeech-HR – a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In *Proceedings of the Third ParlaCLARIN Workshop*, Marseille, France.

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

Machálek, T. (2020). KonText: Advanced and Flexible Corpus Query Interface. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7003–7008, Marseille, France, May. European Language Resources Association. https://www.aclweb.org/anthology/2020.lrec-1.865.

Rayson, P., Dawn Archer, S. P., and McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of the Workshop "Beyond Named Entity Recognition – Semantic labelling for NLP tasks*, pages 7–12. https://eprints.lancs.ac.uk/id/eprint/1783/.

Skubic, J. and Fišer, D. (2022). Parliamentary discourse research in sociology: Literature review. In *Proceedings of the Third ParlaCLARIN Workshop*, Marseille, France.