The Role of Language Technologies in Digital Humanities (The Case of Parliamentary Debates)

Petya Osenova[0000-0002-4484-5027]

Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,
2, Georgi Bonchev Str., Sofia, 1113, Bulgaria
petya@bultreebank.org

Abstract. The paper focuses on the use case of parliamentary debates as part of Digital Humanities. First, the ParlaMint project is outlined as a flagship initiative of CLARIN ERIC infrastructure. The project makes content from the national and regional parliaments visible, comparable and accessible for policy making and research. Then, the approaches are considered that have been applied in the creation of 31 corpora from national and regional parliaments. Last but not least, the utility of the multilingual resource is discussed.

Keywords: Parliamentary Debates, Parlamint, Comparable Corpora, Language Technology, Digital Humanities.

1 Introduction

Digital Humanities (DH) refer to a relatively recent interdisciplinary scientific area in which computational methods are used for conducting deeper studies in literature, history, culture, anthropology, political and social studies, arts, etc. These computational methods facilitate various activities not only in academic institutions but also in administrative ones as well as in museums, libraries, galleries. The activities include collecting and analysing huge data, making content searchable and extractable, visualizing content, observing data in aggregate and statistical ways, etc.

Language technology (LT) and Natural Language Processing (NLP) belong to the computational methods that aim to facilitate studies in DH since they handle texts and speech in both modes – analytical and generative. LT and NLP help with linguistic data processing. They also provide techniques for multimodal applications. Even in the era of powerful transformers like BERT (Devlin, 2019), RoBERTa (Liu, 2019) and AI-induced services like ChatGPT, the collecting, standardizing and processing of text data by language knowledge aware models is more than necessary. The reasons for this are manyfold. For example, data are very often scarce and/or biased; each domain requires special attention due to representation specifics; not all needed information can be learnt only from raw data since there are idiosyncratic and difficult for modelling phenomena.

Digital Presentation and Preservation of Cultural and Scientific Heritage. Conference Proceedings. Vol. 13, Sofia, Bulgaria: Institute of Mathematics and Informatics – BAS, 2023. ISSN: 1314-4006, eISSN: 2535-0366

In this paper I focus on the use case of the Parliamentary debates. I describe how LT and NLP help in preparing parliamentary data for nearly 31 national and reginal parliaments in such a way that they can be used by researchers from political and social studies, linguists, policy makers, any interested parties.

2 The ParlaMint Project: An Overview

ParlaMint (ParlaMint, n.d.) is a flagship project of the research infrastructure for language as social and cultural data CLARIN ERIC (Infrastructure, n.d.). This project contributes to the creation of comparable and uniformly annotated multilingual corpora of parliamentary sessions. It is being conducted in two stages. ParlaMint I was performed in the period of July 2020 – May 2021, while ParlaMint II started in December 2021 and will end in September 2023 (see the list of contributors (ParlaMint-Partners, n.d.)).

The pilot phase ParlaMint I created and made available corpora for 17 parliaments (Erjavec, 2023). In this phase a comparable XML schema in the TEI standard was initiated that allowed for the encoding of all differing parliamentary data. This step built on the previously developed ParlaCLARIN schema (ParlaCLARIN, n.d.). Then, a specialized ParlaMint schema (ParlaMint-Schema, n.d.) was created. Also, rich metadata was added such as information about each member of the respective parliament, about each party, about reactions during parliament sessions (applauses, shouting, laughing, leaving, etc.); an initial validation workflow was established. Already in this phase the corpora started to be used for training and research.

In the subsequent phase - ParlaMint II – another set of 14 parliamentary corpora were added, among which also regional ones (Catalonian, Galician, Basque). During this period the XML schema was improved; the communication among contributors was organized exclusively through GitHub due to the exponential quantities of incoming data; the data of the initial members was extended to cover 2022 and partly 2023; additional metadata was added such as ministers and political orientations; factorised common taxonomies with translations; better encodings of particular phenomena.

In Fig. 1 the metadata information is given for the Bulgarian parliamentary group *Продължаваме промяната* (*We continue the change*). It includes the name with its abbreviation in Bulgarian and English, the date of its establishment and links to its profiles in Wikipedia.

In such a way all the metadata is encoded for parties, governments and other related institutions within the respective time span of the parliament. The challenges here come from the dynamics of newly emerging groups and transformations from groups to parties, etc.

In this version of corpora, the work of specialized committees in parliaments, or the results from voting are not presented for the sake of achieving a comparable resource with the sessions. However, the schema allows for such additions in future.

This example reflects a young parliamentary group in Bulgaria but when it encodes a party with a longer history, then the information about changing names, etc. should be also recorded.

```
<org role="parliamentaryGroup"
     xml:id="parliamentaryGroup.WCC"
     xml:lang="bg">
   <orgName full="yes" xml:lang="bg">Парламентарна група: Продължаваме Промяната</orgName>
   <orgName full="yes" xml:lang="en">Parliamentary Group: We Continue the Change</orgName>
   <orgName full="abb">WCC</orgName>
   <event from="2021-12-03">
      <label xml:lang="en">existence</label>
   </event>
   <idno subtype="wikimedia" type="URI" xml:lang="bg">https://bg.wikipedia.org/wiki/Продължаваме_промяната</idno>
   <idno subtype="wikimedia" type="URI" xml:lang="en">https://en.wikipedia.org/wiki/We_Continue_the_Change</idno>
</org>
```

Fig. 1. Here an XML excerpt from the metadata file about parties and organizations is given out of the Bulgarian corpus header.

At the moment corpora of parliamentary sessions have been created about the following countries or regions: Austria, Basque Country, Bosnia and Herzegovina, Belgium, Bulgaria, Catalonia, Croatia, Czechia, Denmark, Estonia, Finland, France, Galicia, Greece, Hungary, Iceland, Italy, Latvia, Lithuania, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovenia, Spain, Sweden, Turkey, the UK, and Ukraine.

The rationale behind the ParlaMint project are as follows: the EU Parliament has been processed a lot, and its content has been made available through various channels as data (DCEP, n.d.). However, the analysing content from the national parliaments has never been a task of its own since it is not trivial to ensure the comparability among these corpora. At the same time national and regional parliaments contain valuable data that should be taken into account in policy making within their countries as well as within Europe and the world. In addition, these data would be valuable also to DH as material for observing political and social tendencies in the society, and thus facilitating the detection of any societal alerts in time.

The goal of the ParlaMint project is to turn the existing contemporary diverse national parliamentary data into resources that are visible, comparable, comprehensible and accessible to interested parties. A task for the future is the linking of these data with the EU Parliament content.

3 The Workflow in ParlaMint

Here I focus on two types of information.

In section 3.1 the process of corpora creation, processing, validation and delivery is described in more detail. All the steps of the workflow are outlined.

In section 3.2 the tasks envisaged in ParlaMint II are presented. Some of them continue from ParlaMint I while others are newly added.

In fact, ParlaMint II builds on the previous phase and reflects the lessons learnt up to now as well as adds new dimensions with respect to the corpus utility.

3.1 The Process of Corpora Creation

All the contributors had to follow some common steps. These are: a) downloading the parliamentary sessions for their national or regional parliament; b) converting the sessions into a dedicated TEI-based XML schema of ParlaMint, as presented above; c) collecting the respective metadata about the members of the parliament, the speakers in parliament and the parties/parliamentary groups; d) linguistically annotating corpora; e) providing corpora samples via GitHub to the responsible colleagues and reflecting their feedback in the header and all data until achieving error-free data.

Concerning a), the project aimed at covering data from 2015 to 2022 without imposing any restriction on adding available data before 2015 or beyond 2022. Since various corpora had different starting points and approaches, they varied in their time spans. However, for the sake of comparability, a common period had to be set. Also, the data was scraped in various ways – from the parliament site through API, or as HTML, or directly from the parliament internal sources.

As for b) again various techniques were used – automatic or semi-automatic conversions have been performed to have the data in the required ParlaMint format. It had to be taken into account by the schema that two types of parliaments are considered – unicameral and bicameral.

The challenge in c) was the fact that very often the metadata to be gathered was not available on the same place or was completely missing on internet sources (for example, for some guest speakers, metadata could be hardly found). This is a problem especially when there were pre-term elections and new parliaments were often elected, or when many parties formed the parliament ruling majority and respectively – the opposition; also when parties changed their status or the status was not quite clear.

The process in d) requires the availability of an NLP pipeline for the languages which includes tokenization, morphological analysis, syntactic analysis and named entities detection. The output of this pipe should be in the format of Universal Dependencies (UD, n.d.) as the current benchmark in NLP. Most corpora used Stanza models of Stanford (Stanford, n.d.).

Concerning e), a special workflow was designed on GitHub for the contributors to upload a corpus sample, receive feedback and debug errors. This step might take some time especially in case of errors with regard to missing or wrong metadata.

3.2 Tasks in ParlaMint II

In this phase of the project the following work packages (WP) are envisaged (ParlaMint-II, n.d.):

- **WP1:** Documentation, Interoperability, Metadata
- **WP2:** Corpus Expansion
- **WP3:** Corpus Enrichment
- **WP4:** Engagement Activities
- **WP5:** Coordination

WP1 relates to further adjustments and improvements of the ParlaMint encoding schema and GitHub workflow organization.

WP2 takes care of the corpora extension from ParlaMint I to year 2022 and partly to year 2023, as well as of the addition of new parliamentary corpora. The more corpora are added, the more stable the ParlaMint schema becomes for future usages by new contributors.

WP3 introduces an already new dimension of the multilingual corpus utility since it ensures translations of all corpora into English, thus making them comparable de facto. Nowadays the machine translation models are providing good enough quality. Several Machine Translation models were compared among which DeepL (DeepL, n.d.), OPUS-MT (Opus-MT, n.d.), Google Translate (Google, n.d.), Facebook's models mBART50 (Tang, 2020), M2M100 (Fan, 2020). The responsible task leaders decided to use OPUS-MT since it is freely available, easy to use through EasyNMT, supports all involved languages and after some proper name correction gets comparable performance to DeepL.

When translated into English, the corpora are also being semantically tagged. This means that they get annotations of word meanings in context. In this way their content can be also semantically compared across parliaments. For the semantic tagging the state-of-the-art UCREL Semantic Analysis System (USAS) has been used (USAS, n.d.). In addition, multimodal models have been trained and tested on a selected set of parliamentary corpora like the Czech, Polish and Croatian ones (ParlaMintSpeech, n.d.). The ultimate goal even beyond this project would be the output of multimodal data, i.e., the alignment of texts to speeches and video recordings of the sessions.

WP4 is devoted to engagement activities which explore ParlaMint corpora. Such activities are: the creation and delivery of tutorials (Fišer, 2021), participation in hackathons (DHH23, n.d.) and shared tasks, enhancing various showcases such as the impact stories. One of these stories features gender analysis in selected parliaments (Skubic, 2023). Another one focuses on discourse around migration during 2015/16 and 2020 (Del Fante, 2023). There is also an impact story that surveyed the topics during COVID-19 period (Calabretta, 2021).

WP5 is handling the project organization and is seeking for more paths to go with expanding, exploring and linking parliamentary data.

4 Exploitation of ParlaMint Corpora

The corpora are available in two modes: as data (Erjavec T. e., 2021)and in concordancing tools. At the moment version 2.1 are available but version 3.0 is expected at the end of June 2023 and version 3.1 – at the end of September 2023.

As mentioned before, the corpora as data are available in two versions – only with metadata, and with linguistic annotation. These versions are usable for training language models that are specific to the parliamentary data. For the end users the corpora have been uploaded into NoSketch engine concordancing tool (NoSketchEngine) where the data and metadata can be observed in context; can be visualized accordingly; and various implemented statistics can be applied over data and metadata.

In Fig. 2 an excerpt from the concordance of the lemma безработица (unemployment) is presented. On the left, the name of the member of parliament is

shown who authors the utterance. Also, the date when the utterance was delivered is available.

GyokovGeorgi,2018-02-23	! Уважаеми господин Министър, дългосрочната	безработица	или така наречените „продължително
GyokovGeorgi,2018-02-23	е първостепенна задача, защото дългосрочната	безработица	е проблем не само за самите безработн
GyokovGeorgi,2018-02-23	на Стратегия „Европа 2020“, и за намаляване на	безработицата	като цяло. ∮ Затова е важно: какви дейс
PetkovBiser,2018-02-23	, политиката за намаляване на продължителната	безработица	е приоритет в работата на Министерств
PetkovBiser,2018-02-23	работа и съкращаване на продължителността на	безработицата	. При предоставянето на услугите за акт
PetkovBiser,2018-02-23	на анализ на причините за дългия период на	безработица	и личните потребности на лицето се пре
PetkovBiser,2018-02-23	безработни лица от общини с високо равнище на	безработица	възстановяват или придобиват трудови
PetkovBiser,2018-02-23	в цялата страна, в 193 общини с равнище на	безработица	над средното по Националната схема „F
GyokovGeorgi,2018-02-23	с всичките си действия и политики по заетостта и	безработицата	си затваря очите, когато работодателит
VeselinovIskren,2014-12-09	парадират, че са техни, хората са осъдени на	безработица	и мизерия. Нищо не е направено за тези
AdemovHasan,2014-12-09	и само на ръста на средния осигурителен доход?	Безработицата	я планирате на същите нива - 11,7. ∮ От

Fig. 2. Here an excerpt from the Bulgarian corpus is shown about the usage of lemma безработица (unemployment).

Apart from the debates in parliamentary sessions, the accompanying reactions are also very valuable to observe. In Fig. 3 the statistics of these reactions is shown for the Bulgarian parliament.

It can be seen that the most frequent reaction is *interruption*. Then comes a*pplause* followed by *signal* which indicated the wish of the member of the Parliament to be given the word. Next is the *noise* in the plenary room. After noise the more emotionally weighted reactions come – *exclamation* and *laughter*.

It would be interesting to investigate when the certain reaction is applied across parliaments. For example, applause in the Bulgarian parliament is related mostly to support of the members of the same parliamentary group to a speaker from the same group or to official speeches given by the President of the state, diplomats from EU, etc.

What is not visible in Fig. 3 are the following not so frequent phenomena: *shouting, leaving, greeting, entering, question, clarification.*

Corpus: **ParlaMint-BG 2.1 (Bulgarian parliament)**
Total number of items: **19**
Total frequency: **34,138**

note.type	document frequency
vocal:interruption	16,155
kinesic:applause	7,625
kinesic:signal	2,221
vocal:noise	1,711
vocal:exclamat	1,598
vocal:laughter	1,289
kinesic:ringing	1,175
vocal:speaking	735
incident:action	644

Fig. 3. Here an excerpt from the Bulgarian corpus is shown about the frequency of accompanying reactions.

The data is searchable by various metadata options. These are: the term of the parliament; member or non-member of parliament; chairperson or regular speaker; speaker's party; speaker's party status – coalition or opposition; name, gender, birthday of speaker.

5 Conclusions

The multilingual and multi-parliamentary ParlaMint project was started by CLARIN ERIC at times of health crisis - the outbreak of COVID-19. The idea behind this project is to pave the way of making the national and regional corpora visible, comparable and usable also with respect to periods of global disturbances. The project continues in times of another crisis - social and political – Russia's war against Ukraine. These two major crises marked the division of parliamentary corpora in three subsets: reference, covid-aware and war-aware.

ParlaMint is more than a project – it has implemented a whole infrastructure and life cycle for adding and processing further parliamentary data – with starting from unifying formats, adhering to the same standard; going with adding metadata and linguistic processing; ensuring validation and consistency, and finally making data publicly available for training and research. The parliamentary corpora are invaluable resource to DH because they provide observable data with rich metadata.

Acknowledgements.

The reported work has been partially supported by CLaDA-BG, the Bulgarian National Interdisciplinary Research e-Infrastructure for Resources and Technologies in favor of the Bulgarian Language and Cultural Heritage, part of the EU infrastructures CLARIN and DARIAH, Grant number DO1-301/17.12.21.

References

Calabretta, I. a. (2021). *Helsinki Digital Humanities Hackathon 2021: 'Parliamentary Debates in COVID Times'*. https://www.clarin.eu/impact-stories/helsinki-digital-humanities-hackathon-2021-parliamentary-debates-covid-times

DCEP. (n.d.). *Digital Corpus of the European Parliament*. https://joint-research-centre.ec.europa.eu/language-technology-resources/dcep-digital-corpus-european-parliament_en

DeepL. (n.d.). *DeepL translator*. https://www.deepl.com/translator

Del Fante, D. a. (2023). *ParlaMint – A Resource for Democracy*. https://www.clarin.eu/impact-stories/parlamint-resource-democracy

Devlin, J. a.-W. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv:1810.04805

DHH23. (n.d.). *Political polarization*. https://www.helsinki.fi/en/digital-humanities/dhh23-hackathon/dhh23-themes

Erjavec, T. a. (2023). The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation, 57*, 415–448. https://doi.org/10.1007/s10579-021-09574-0

Erjavec, T. e. (2021). *Multilingual comparable corpora of parliamentary debates ParlaMint 2.1*. CLARIN ERIC. https://doi.org/10.1007/s10579-021-09574-0

Fan, A. a.-K. (2020). *Beyond English-Centric Multilingual Machine Translation*. https://arxiv.org/abs/2010.11125
Fišer, D. a. (2021). *Voices of the Parliament: A Corpus Approach to Parliamentary Discourse Research.* Institute of Contemporary History. https://sidih.si/cdn/121/index.html
Google. (n.d.). *Google Translate*. https://pypi.org/project/googletrans/
Infrastructure, C. L. (n.d.). *CLARIN*. https://www.clarin.eu/
Liu, Y. a. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach.* arXiv:1907.11692
NoSketchEngine. (n.d.). ParliaMint corpora. https://www.clarin.si/noske/
Opus-MT. (n.d.). *Opus-MT*. https://github.com/Helsinki-NLP/Opus-MT
ParlaCLARIN. (n.d.). *Parla-CLARIN Schema*. https://github.com/clarin-eric/parla-clarin
ParlaMint. (n.d.). *ParlaMint: Towards Comparable Parliamentary Corpora*. https://www.clarin.eu/parlamint
ParlaMint-II. (n.d.). *ParlaMint II*. https://www.clarin.eu/parlamint#parlamint-ii
ParlaMint-Partners. (n.d.). *Project Partners*. https://www.clarin.eu/parlamint#Partners
ParlaMint-Schema. (n.d.). *ParlaMint-Schema*. https://github.com/clarin-eric/ParlaMint/blob/main/Schema/README.md
ParlaMintSpeech. (n.d.). *ASR training dataset for Croatian ParlaSpeech-HR v1.0*. https://www.clarin.si/repository/xmlui/handle/11356/1494
Skubic, J. a. (2023). *Networks of Power - Gender Analysis in European Parliaments*. https://www.clarin.eu/impact-stories/networks-power-gender-analysis-european-parliaments
Stanford. (n.d.). *Stanza*. https://stanfordnlp.github.io/stanza/available_models.html
Tang, Y. a.-J. (2020). *Multilingual Translation with Extensible Multilingual Pretraining and Finetuning.* https://arxiv.org/abs/2008.00401
UD. (n.d.). *Universal Dependencies*. https://universaldependencies.org/
USAS. (n.d.). *UCREL Semantic Analysis System (USAS)*. http://ucrel.lancs.ac.uk/usas/

Received: May 20, 2023
Reviewed: May 27, 2023
Finally Accepted: June 15, 2023