Guidelines for adding new corpora to ParlaMint

Authors: Maciej Ogrodniczuk, Petya Osenova, Tomaž Erjavec

Last change: 2020-10-26

1 Introduction

This document describes the process of adding new corpora to ParlaMint, i.e. the process of collecting, processing and showcasing national parliamentary data.

2 Gathering the data

While we expect that most applicants will already have experience in gathering the data, or have in fact already done so, for those that have not, here is a brief overview of the process.

The data can be gathered automatically by a dedicated crawler from the parliament website, or, better, obtained as a database dump from the Parliament directly. The speaker metadata will probably have to be downloaded separately, maybe from several sources. Thus, our advice is first to make some research on how much time you will need to develop the scripts to download the data and metadata.

Time span

The data should cover the period between 2015 and 2020 (or an equivalent one, preferably full cadencies).

The corpora will be divided into two parts:

- COVID-19 Parliament Corpus (Oct 2019 - July 2020 (and later, if possible)
- Reference Parliamentary Corpus (~2015 - Oct 2019)

Data types and characteristics

The data in question are textual data representing at least session transcripts, split into speeches, with accompanying speech and speaker metadata as explained below. It should be noted that the speaker metadata is split between *static metadata* (which does not change), and *dynamic metadata*, which can change through time, and is therefore ideally time-stamped.

Speech data

The transcriptions need to be split into mandates, sessions and speeches, with appropriate meta-data given to each. Each speech can, minimally, contain only pure transcripts of the speeches. However, many transcripts also contain commentary by the transcribers, such as */microphone turned off/, /unclear/, /clapping/*, etc. These then need to be converted into the proper Parla-CLARIN encoding.

Speaker metadata

The following static metadata **must** be available for each speaker:

- speaker's ID
- name and surname(s)
- gender

The following static metadata **may** be available for a speaker:

- birth date
- birth place
- death date
- death place
- education
- employment (although in principle dynamic, this information is typically available only for the employment when the person becomes an MP)
- link to Wikipedia, VIAF or other authoritative records

The following dynamic metadata **must** be available for each speaker:

- status of the speaker (MP, invited speaker)
- political affiliation (only necessary for MPs, not for occasional speakers)

The political affiliation gives only a reference to the ID of the party or parties, the information on which is stored separately.

The following information **must** be given for each political party:

- abbreviation of the party
- full name of the party in the local language,
- the party ID (potentially the same as the abbreviation).

The following information **may** be given for a political party:

- abbreviation of the party in English,
- full name of the party in English,
- time-stamped change of names (abbreviations, full names) of the party
- relation to other parties

In case other speaker metadata is available, it may also be encoded in the corpus.

Speech metadata

Each speech also has associated metadata, either directly, or inherited from superordinate elements (e.g. if the date of the session is known, then so is the date of the contained speech).

The following metadata **must** be available for each speech:

- date of the speech (but can be approximate, if exact date is unknown)
- speaker ID
- role of the speaker (one of: chair (required), regular (required), irregular (optional)).

The following metadata **may** be available for a speech:

- time of speech
- type of speech (question, answer)
- to whom the speech is addressed
- summary of the speech

In case other speech metadata is available, it may also be encoded in the corpus.

Linguistic annotation

The base corpus must also be linguistically annotated, and the annotated corpus stored separately. The linguistic annotation **must** contain tokenisation, sentence segmentation, lemmatisation, annotation with UD part-of-speech and morphological features, UD dependency relations and named entity markup (PER, ORG, LOC, MISC).

These annotations might be created with an in-house UD and NER pipeline, or the UD annotation can be performed with UD-Pipe or other tools trained on UD treebanks.

3 Converting the data to Parla-CLARIN format

The Parla-CLARIN format is a parameterization of the TEI Guidelines created for scholarly encoding of parliamentary corpora. Please note that in ParlaMint we use this format that is further specified. It is maintained on the GitHub project https://github.com/clarin-eric/parla-clarin/, which contains a file with the TEI parameterization and the documentation (i.e. a TEI ODD XML document), the generated XML schema in various schema languages (RelaxNG, W3C schema, DTD), as well as complete example documents from several projects. The ODD documentation is also available for reading (i.e. converted to HTML) at https://clarin-eric.github.io/parla-clarin/. Please read it before starting to encode the corpus.

However, the most precise and up-to-date guide to the ParlaMint encoding is in fact the CLARIN repository entry http://hdl.handle.net/11356/1345. The entry contains the already finished corpora which can serve as exemplars to what is encoded and how, and the entry also includes RelaxNG schemas developed esp. for ParlaMint, which specialise the more general Parla-CLARIN schema. These schemas should be used to validate the encoding of ParlaMint corpora.

Structure of the corpus

The general structure of the corpus is explained in Section 3. "Overall document structure". For ParlaMint, we make the following further requirements:

- The complete corpus for country XX is stored in a directory called `ParlaMint-XX/`
- The directory should contain the root file of the corpus named `ParlaMint-XX.xml`. This file has the root XML element `<teiCorpus>`, followed by the corpus `<teiHeader>` giving the meta-data of the corpus, followed by XInclude directives of the component files.
- The directory should also contain the component files of the corpus, with the names of the form `ParlaMint-XX_<suffix>.xml`. The `<suffix>` and what the scope of one component file is, will depend on the original data, but it is often one session

on a particular date, e.g. `ParlaMint-XX_2020-06-25.xml`. or, if there are more sessions on the date e.g. `ParlaMint-XX_2020-06-25-1.xml`. The sort order of the files should be chronological.

- The corpus exists in two variants, the one with the structured metadata and data, as above, and the one with linguistically analysed text content of the <seg> elements. This linguistically analysed corpus should be stored in the directory `ParlaMint-XX.ana/` which contains files `ParlaMint-XX.ana.xml` and `ParlaMint-XX_<suffix>.ana.xml`. It is not necessary to change the IDs inside the corpus files.

Corpus header

The corpus header for the 4 countries can be viewed through the file ParlaMint-XX.xml where XX replaces the country name - BG, HR, PL and SI. Here only some special points will be mentioned. The full information can be obtained from the TEI headers of the current ParlaMint corpora and in the dedicated descriptions of the used Parla-CLARIN format.

First, the **metadata about the creators** is given:

<respStmt>

<persName xml:lang="bg">Владислава Григорова**</persName>**

<persName xml:lang="en">Vladislava Grigorova**</persName>**

<resp xml:lang="bg">Сваляне, почистване и конвертиране до XML**</resp>**

<resp xml:lang="en">Download, clean-up converting of the data to XML**</resp>**

</respStmt>

Then, **types of meetings** are described:

<catDesc xml:lang="en"><term>Types of meetings**</term></catDesc>**

<category xml:id="parla.meeting.regular">

<catDesc xml:lang="en"><term>Regular meeting**</term></catDesc>**

<catDesc xml:lang="bg"><term>Редовно заседание**</term></catDesc>**

</category>

<category xml:id="parla.meeting.special">

<catDesc xml:lang="en"><term>Special meeting**</term></catDesc>**

<catDesc xml:lang="bg"><term>Специално заседание**</term></catDesc>**

<category xml:id="parla.meeting.extraordinary">

<catDesc xml:lang="en"><term>Extraordinary meeting**</term></catDesc>**

<catDesc xml:lang="bg"><term>Извънредно заседание**</term></catDesc>**

</category> [...]

Also, **types of speakers** are listed:

<taxonomy xml:id="speaker_types">

<desc xml:lang="en"><term>Types of speakers**</term></desc>**

<desc xml:lang="bg"><term>Видове изказващи се**</term></desc>**

<category xml:id="chair">

<catDesc xml:lang="bg"><term>Председател**</term>**: председател на заседание**</catDesc>**

<catDesc xml:lang="en"><term>Chairperson**</term>**: chairman of a meeting**</catDesc>**

</category>

<category xml:id="regular">

<catDesc xml:lang="bg"><term>Говорещ**</term>**: обикновен изказващ се на заседание**</catDesc>**

<catDesc xml:lang="en"><term>Regular**</term>**: a regular speaker at a meeting**</catDesc>**

</category>

<category xml:id="guest">

<catDesc xml:lang="bg"><term>Гост**</term>**: гост на заседание</catDesc>

<catDesc xml:lang="en"><term>Guest**</term>**: a guest speaker at a meeting**</catDesc>**

</category>

</taxonomy>

From the metadata there is a **list of parties**:

<listOrg>

<head xml:lang="bg">Партии в Народно събрание**</head>**

<**head xml:lang="en">**Parties in Bulgarian parliament**</head>**

<org role="political_party" xml:id="party.GERB">

<orgName full="yes" xml:lang="bg">Граждани за европейско развитие на България**</orgName>**

<orgName full="init" xml:lang="en">GERB**</orgName>**

<orgName full="init">ГЕРБ**</orgName>**

<event from="2006-12-03">

```
                    <label xml:lang="en">existence</label>
                </event>
            <idno type="wikimedia" xml:lang="bg">
                https://bg.wikipedia.org/wiki/ГЕРБ</idno>
            <idno type="wikimedia" xml:lang="en">
                https://en.wikipedia.org/wiki/GERB</idno>
            </org>

        <org role="coalition" xml:id="party.BSPBG">
            <orgName full="yes" xml:lang="bg">БСП за България</orgName>
            <orgName full="yes" xml:lang="en">Coalition for Bulgaria</orgName>
                <orgName full="init">БСП</orgName>
                <event from="2017">
                    <label xml:lang="en">existence</label>
                </event>
            <idno type="wikimedia" xml:lang="bg">
                https://bg.wikipedia.org/wiki/БСП_за_България</idno>
            <idno type="wikimedia" xml:lang="en">
                https://en.wikipedia.org/wiki/Coalition_for_Bulgaria</idno>
                </org>
```

Another important type of metadata is the **list of Persons:**

```
<listPerson>
        <head xml:lang="bg">Депутати</head>
        <head xml:lang="en">Members of Parliament</head>
            <person n="2801" xml:id="ShevkedAdlen">
                <persName>
                    <forename>Адлен</forename>
                    <surname>Шукри</surname>
                    <surname>Шевкед</surname>
                </persName>
                <sex value="F">женски</sex>
                <birth when="1969-11-24">
                    <placeName>Кърджали, България</placeName>
```

```
            </birth>
             <affiliation ref="#NS" role="MP">депутат</affiliation>
            <affiliation ref="#party.DPS" role="member"/>
        </person>
```

The **list of used languages** is added:

```
 <langUsage>
        <language ident="bg" xml:lang="bg">български</language>
        <language ident="bg" xml:lang="en">Bulgarian</language>
        <language ident="en" xml:lang="bg">английски</language>
        <language ident="en" xml:lang="en">English</language>
        <language ident="fr" xml:lang="bg">френски</language>
        <language ident="fr" xml:lang="en">French</language>
 </langUsage>
```

Finally, the **data file**s are specified through *xi: include*:

```
<xi:include xmlns:xi="http://www.w3.org/2001/XInclude"
href="ParlaMint-BG_2014-10-27.xml"/>
  <xi:include xmlns:xi="http://www.w3.org/2001/XInclude"
href="ParlaMint-BG_2014-10-29.xml"/>
  <xi:include xmlns:xi="http://www.w3.org/2001/XInclude"
href="ParlaMint-BG_2014-10-30.xml"/>
  <xi:include xmlns:xi="http://www.w3.org/2001/XInclude"
href="ParlaMint-BG_2014-10-31.xml"/>
```

4 Describing the corpus

In the last step, the corpus should be documented by adding a short description of the data, information about the source, markup process, structure and statistics to the document listing all ParlaMint corpora. Information about the date and name of the author of the description should be also provided in the table in the document header.

Here is an example description:

Basic statistics of the subcorpora

Language	Time span	Number of utterances	Number of words
Bulgarian	2014-10-27 – 2020-07-31	147 432	18 757 284
Polish	2015-11-12 – 2020-08-18	331 082	27 448 141

1 Polish Corpus

Source and conversion of data

The data and linguistic annotation was retrieved from The Polish Parliamentary Corpus (http://clip.ipipan.waw.pl/PPC) and converted to Parla-CLARIN format from its internal TEI P5 XML representation following the format of the National Corpus of Polish (http://www.nkjp.pl). The conversion was performed with a set of Python scripts. Some errors in the original corpus were automatically corrected during conversion.

Corpus statistics

The corpus contains the stenographic record of plenary sittings of the Sejm (8th and 9th term of office) and Senate (9th and 10th term of office) of the Republic of Poland. It is composed of two subcorpora: the reference subcorpus, with utterances between 2015-11-12 and 2019-10-31 and COVID subcorpus, between 2019-11-01 and 2020-08-18. Both subcorpora contain 516 files representing individual session days, 330k utterances and 27M words (additional statistics are available at the NoSketch Engine corpus info page):

	Sejm		Senate	
Term of office	8 (2015–2019)	9 (2019–2020)	9 (2015–2019)	10 (2019–2020)
Session days	235	43	204	34
	278		**238**	
Number of utterances	182 050	24 406	110 760	13 866
	206 456		**124 626**	
				331 082
Number of	15 864 348	1 858 701	8 377 720	1 347 372

words	**17 723 049**	**9 725 092**
		27 448 141

Corpus metadata

MP metadata (gender, birth date, political affiliations) were retrieved from the parliament website. The speakers were assigned a role of chairman, regular or guest. All MPs were given the role of regular, even if they are speaking as PM, minister or someone else.

Corpus annotation

Named entities were converted from PPC to Parla-CLARIN: time and date entities were ignored, as they are not recognised by Parla-CLARIN specifications. Subtypes were skipped.

Heuristics was used to convert comments into Parla-CLARIN incidents, mostly based on typical phrases used in comment text.

Notes were assigned Parla-CLARIN types, also based on typical phrases.