ParlaMint-NO

Norwegian ParlaMint data.

This corpus contains the Norwegian part of the ParlaMint project. It contains minutes from the following meetings:

- Meetings in Stortinget, October 1998 - May 2022
- Meetings in Lagtinget and Odelstinget, October 1998 - June 2009

Text data is taken from https://data.stortinget.no/, and converted to ParlaMint with XSLT and Python. Metadata for Stortinget representatives and ministers is taken from Stortinget's API and Wikidata.

ParlaMint

ParlaMint is an EU-funded project supported by CLARIN ERIC. CLARIN stands for *Common Language Resources and Technology Infrastructure*, while ERIC stands for *European Research Infrastructure Consortium*. The project's aim is to create comparable and similarly annotated corpora of parliamentary meeting minutes.

- Project page: https://www.clarin.eu/parlamint.
- Github: https://github.com/clarin-eric/ParlaMint/.

Format and structure

The corpus consists of xml files in ParlaMint format, a specialization of Parla-CLARIN. Parla-CLARIN is a TEI format designed to describe parliamentary minutes. The ParlaMint schema is described in detail in the project's GitHub repository. The more general Parla-CLARIN is described here.

The corpus consists of two parts: "ParlaMint-NO.TEI" consists of Stortinget minutes with information on speakers. "ParlaMint-NO.TEI.ana" contains the same texts annotated at word and sentence level. The automatic linguistic analysis was carried out with the Python library Spacy and a self-adapted model for both Bokmål and Nynorsk based on *Norsk dependenstrebank*, *Norwegian Dependency Treebank* (NDT).

Each part of the corpus has a root file (`ParlaMint-NO.xml`, `ParlaMint-NO.ana.xml`) and otherwise consists of component files. The component files are organized in folders by calendar year (NOTE: the folder structure does not follow the parliamentary year). Each component file name consists of the stem "ParlaMint-NO_" and the date of the meeting. Meetings in Lagtinget have "-upper" in the name, while meetings in Odelstinget have "-lower". Some days have more than one meeting. Meetings after the first have a number denoting their order at the end, eg. "-*N*". There

are no days with more than two meetings in the same chamber. Each TEI file (`.xml` extension) has a corresponding TEI.ana file with extension `.ana.xml` based on the same content.

Each part file has a header element (`teiHeader`) that contains metadata about the content, and an element `text` with the file's content.

Root files contain metadata for each part of the corpus, including:

- Number of words and sentences
- Metadata for speakers
- Metadata for organizations
- Taxonomies

Metadata

ParlaMint schema.

Two parts (`TEI, TEI.ana`).

Each part consists of:

- Root file
- 3.267 part files

The corpus contains

- 398.781 utterances
- 97.541.932 words

Language:

- Bokmål (Norwegian)
- Nynorsk (Norwegian)

Resources

The ParlaMint project offers scripts and tools for working with ParlaMint data in the project's GitHub repository. This includes tasks such as:

- all text
- Convert to `.conllu`

Comments

About **Lagtinget** and **Odelstinget**: Since 1814, Norway has formally had a unicameral system, but in practice between 1814 and 2009 it was a variant of a bicameral system. During this period, Stortinget divided itself into two departments, Odelstinget and Lagtinget, which had the functions of lower house and upper house,

respectively. The system with Lagtinget and Odelstinget was used for the last time in 2009.

In ParlaMint-NO, this means that in the period 1998-2009 there are three types of minutes:

- From Odelstinget, marked with "-lower" in the file title.
- From Lagtinget, marked with "-upper" in the file title.
- From both departments, without any marking.

From 2009 on, these distinctions are not relevant.

MP's language form: The transcripts are in one of the two language forms of written Norwegian, Bokmål or Nynorsk. Members have the option to choose which language form they will be referred to in.

License

Creative Commons-ZERO (CC-ZERO)

Links

Stortinget API: https://data.stortinget.no/

ParlaMint II project page: https://www.clarin.eu/parlamint

ParlaMint - Github: https://github.com/clarin-eric/ParlaMint /

Clarin ERIC: https://www.clarin.eu/

Github repo with scripts to create this corpus:
https://github.com/tungland/ParlaMint-NO-scripts