ParlaMint-NO

Norske ParlaMint-data.

Dette korpuset inneholder den norske delen av ParlaMint-prosjektet. Det inneholder referater fra følgende møter:

- Møter i Stortinget oktober: 1998 - mai 2022
- Lagtinget og Odelstinget oktober: 1998 - juni 2009

Tekstdata er hentet fra https://data.stortinget.no/, og konvertert til ParlaMint med XSLT og Python. Metadata for stortingsrepresentanter og ministre er hentet fra Stortingets API og Wikidata.

ParlaMint

ParlaMint er et EU-finansiert prosjekt støttet av CLARIN ERIC, et europeisk konsortium for forskningsinfrastruktur for språkressurser. Prosjektets mål er å lage sammenlignbare og likt annoterte korpus av parlamentariske møtereferater.

Prosjektside: https://www.clarin.eu/parlamint.

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

Format og struktur

Korpuset består av xml-filer i ParlaMint-format, en spesialisering av Parla-CLARIN. Parla-CLARIN er et TEI-format laget for å beskrive parlamentariske referater. ParlaMint-skjemaet er beskrevet i detalj i prosjektets GitHub-repositorium. Det mer generelle Parla-CLARIN er beskrevet her.

Korpuset består av to deler. Den ene ("ParlaMint-NO.TEI") består av Stortingsreferater med informasjon om talere. Den andre ("ParlaMint-NO.TEI.ana") inneholder de samme tekstene annotert på ord- og setningsnivå. Den automatiske lingvistiske analysen ble utført med Python-biblioteket Spacy og en egentilpasset modell for både bokmål og nynorsk basert på Norsk dependenstrebank (NDT).

Hver del av korpuset har en rot-fil (`ParlaMint-NO.xml`, `ParlaMint-NO.ana.xml`) og består forøvrig av delfiler. Delfilene er sortert i mapper etter kalenderår (OBS: mappestrukturen følger ikke parlamentarisk år). Hvert delfilnavn består av stammen "ParlaMint-NO_" og møtets dato. Møter i Lagtinget har "-upper" i navnet, mens møter i Odelstinget har "-lower". Noen dager har mer en ett møte. Da har møter etter det

første et løpenummer, eg. "-*N*". Det er ingen dager med mer enn to møter i samme kammer. Hver TEI-fil (`.xml`-utvidelse) har en tilsvarende TEI.ana-fil med utvidelse `.ana.xml` basert på det samme innholdet.

Hver delfil har et hode-element (`teiHeader`) som inneholder metadata om innholdet og et element `text` med filens innhold.

Rot-filene inneholder metadata for hver del av korpuset, inkludert:

- Antall ord og setninger
- Metadata for talere
- Metadata for organisasjoner
- Taksonomier

Metadata

Korpuset består av xml-filer markert i ParlaMint-skjema.

To deler (`TEI`, `TEI.ana`).

Hver del består av:

- Rotfil
- 3267 delfiler

Korpuset inneholder

- 398781 taler
- 97541932 ord

Språk

- Nynorsk
- Bokmål

Ressurser

ParlaMint-prosjektet tilbyr skript og verktøy for å jobbe med ParlaMint-data i prosjektets GitHub-repositorium. Dette inkluderer oppgaver som:

- Hente ut all tekst
- Konvertere til `.conllu`

Kommentarer

Lagting og Odelsting: Norge har siden 1814 formelt hatt et ettkammersystem, men i praksis var det mellom 1814 og 2009 en variant av et tokammersystem. I denne perioden delte Stortinget seg selv i to avdelinger, Odelstinget og Lagtinget, som hadde funksjoner som henholdsvis underhus og overhus. Ordningen med Lagting og Odelsting ble brukt for siste gang i 2009.

I ParlaMint-NO vil det si at det i perioden 1998-2009 finnes tre typer referater:

- Fra Odelstinget, markert med "-lower" i filtittel.
- Fra Lagtinget, markert med "-upper" i filtittel.
- Fra begge avdelinger, uten noen markering.

Fra 2009 finnes ikke disse skillene.

Representanters målform: Referatene er stort sett på Bokmål eller Nynorsk. Representantene har mulighet til å velge hvilken målform de vil refereres i.

Lisens

Lenker:

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

ParlaMint II prosjektside: https://www.clarin.eu/parlamint

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

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

Github repo med scripts for å lage dette korpuset.
https://github.com/tungland/ParlaMint-NO-scripts