ENGLISH TEXT OVERLEAF

Nasjonalbiblioteket

Stortingsforhandlinger 1814-2000

Dett korpuset inneholder dokumenter fra Stortinget for perioden 1814 til 2000, fordelt på totalt 2136 bind. Korpuset har en størrelse på 1,5 milliarder løpende ord (tokens).

Format

Korpuset publiseres på formatet JSONL (JSON lines) med et avsnitt per linje. Hver linje har følgende felt: URN, sidenummer, avsnittsnummer og løpende tekst. Teksten er tokenisert med tokenisator fra DH-LAB ved Nasjonalbiblioteket (https://github.com/DH-LAB-NB/DHLAB). Filnavet inneholder deler av URNen, bare legg til prefikset "URN:NBN:no-nb_" for å få en fullverdig URN.

Metadata

Grunnleggende metadata for hvert bind/hver URN finnes i CSV-filen *metadata.csv*. Metadataene er hentet fra Nasjonalbibliotekets metadata-API (Mods), med kallet:

https://api.nb.no/catalog/v1/metadata/<URN>/mods

f.eks.

https://api.nb.no/catalog/v1/metadata/URN:NBN:no-nb_digistorting_2003_part5_vol-c/mods

Mer omfattende metadata finnes på Stortingets sider:

- https://stortinget.no/no/Saker-og-publikasjoner/Stortingsforhandlinger/

Kommentarer

Dette korpuset inneholder publiserte historiske stortingsforhandlinger fra Stortinget. Bindene ble OCR-lest og prosessert ved Nasjonalbiblioteket, og tilgjengeliggjort på nett i 2014, https://www.nb.no/statsmaktene/. Når de nå gjøres tilgjengelig for nedlasting, legges dataene ut slik de foreligger, uten videre korrektur eller redigering. Det er gått mer enn ti år siden materialet ble prosessert og særlig eldre protokoller (før 1950) kan være av dårlig kvalitet.

I 2017 publiserte Språkteknologigruppa ved UiO *Talk of Norway*-korpuset, en samling stortings-forhandlinger fra 1998 til 2016, se: https://github.com/ltgoslo/talk-of-norway.

Nasjonalbiblioteket er en del av det CLARIN-støttede ParlaMint-prosjektet (https://www.clarin.eu/parlamint) og kommer til å publisere et nytt stortingskorpus fra perioden 1998-2022 i slutten av 2022 i samarbeid med europeiske partnere. Korpuset vil muliggjøre sammenlikninger på tvers av europeiske land.

Lisens

2022-09-07, Språkbanken – Nasjonalbiblioteket

National Library of Norway

Norwegian Parliamentary Proceedings 1814-2000

This corpus contains documents from the Norwegian Parliament from 1814 to 2000, spanning 2136 volumes. The corpus comprises 1.5 billion tokens.

Format

The corpus is published in JSONL (JSON lines) format with one paragraph per line. Each line has the following columns: URN, page number, paragraph number, text. The text is tokenized using the tokenizer of the DH-LAB at the National Library of Norway (https://github.com/DH-LAB-NB/DHLAB). The filename contains the URN, just add the prefix "URN:NBN:no-nb_".

Metadata

Basic metadata for each volume / URN can be found in the CSV file *metadata.csv*. The metadata were harvested from the metadata API (Mods API) of the National Library of Norway, using:

https://api.nb.no/catalog/v1/metadata/<URN>/mods

e.g.

https://api.nb.no/catalog/v1/metadata/URN:NBN:no-nb_digistorting_2003_part5_vol-c/mods

More fine-grained metadata can be found at the Norwegian Parliament:
https://stortinget.no/no/Saker-og-publikasjoner/Stortingsforhandlinger/

Remarks

This corpus contains published historical parliamentary proceedings from the Norwegian parliament. The proceedings were OCR'ed and further processed at the National Library of Norway, and made available at https://www.nb.no/statsmaktene/ in 2014. Now, in 2022, the same data is made available for download. It is provided as is. More than ten years have passed since the material was OCR'ed and older proceedings (before 1950) may be of poor quality.

In 2017, the Language Technology Group at the University of Oslo published *Talk of Norway*, a collection of Norwegian parliament speeches from 1998 to 2016, cf. https://github.com/ltgoslo/talk-of-norway.

The National Library of Norway is part of the CLARIN-supported ParlaMint project (https://www.clarin.eu/parlamint) and will publish a Norwegian parliamentary corpus covering the years 1998-2022 by the end of 2022 in cooperation with European partners. The corpus will enable comparison between European parliamentary data.

Licence

2022-09-07, The Language Bank – The National Library of Norway