Nasjonalbiblioteket
National Library of Norway

Stortinget Speech Corpus version 1.0

Table of Contents

Dataset Description

- **Github repositories**: Transcription matching, metadata extraction
- **Paper:** A Large Norwegian Dataset for Weak Supervision ASR
- **Point of Contact:** The Norwegian Language Bank

Dataset Summary

The Stortinget Speech Corpus (SSC) is a 5000+ hours speech dataset for weak supervision ASR created from audio and aligned proceedings text from Stortinget, the Norwegian Parliament. It contains speech segments of up to 30 seconds with transcriptions in Norwegian Bokmål and Norwegian Nynorsk from the official proceedings. The dataset only uses open data and is distributed with a CC0 license.

Languages

The transcriptions in the dataset are in Norwegian Bokmål (nob) or Norwegian Nynorsk (nno).

Dataset Structure

Data Instances

The dataset is distributed as a JSONL file, `ssc_v1_0.jsonl`. Audio files, proceedings files and transcription files (with ASR output) are included in this repository, and there are relative file paths in the JSONL file. Note that only segmented audio files are part of the release. However, MP4 files with the parliamentary videos used when making this

dataset, are freely accessible from the Stortinget video archive. A list with the link to each video file is given in `ssc_v1_0_video_urls.txt`, which is part of the release.

Example of a segment:


```
{'segment_id': 0,
 'audio_path': 'data/audio/2010/stortinget-20100106-090925_3240100_3267900.mp3',
 'context_after': 'De annonserte statsrådene er til stede, og vi er klare til å starte
den muntlige spørretimen. De representanter som i tillegg til de forhåndspåmeldte ønsker
å stille hovedspørsmål, bes om å reise seg. Vi starter da med første hovedspørsmål, fra
representanten Hans Frode Kielland Asmyhr. Jeg har et spørsmål til statsråd Storberget.
La meg også ønske statsråden et riktig godt nytt år. La oss også håpe at dette ikke blir
et veldig trivelig år for',
 'context_before': 'å ønske alle representanter og andre som har sitt daglige arbeid i
Stortinget, et riktig godt nytt år, og håper at det blir et godt arbeidsår for
Stortinget. Representantene Erling Sande og Line Henriette Hjemdal, som har vært
permittert, har igjen tatt sete. Den',
 'duration': 27.8,
 'meeting_date': '2010-01-06',
 'num_speakers': 1,
 'proceedings_text': 'innkalte vararepresentant for Buskerud fylke, Elizabeth Skogrand,
har tatt sete. Stortinget mottok mandag meddelelse fra Statsministerens kontor om at
utenriksminister Jonas Gahr Støre og statsrådene Knut Storberget og Lars Peder Brekk vil
møte til muntlig spørretime.',
 'proceedingsfile': 'data/proceedings/2010-01-06_proceedings.txt',
 'score': 0.7764705882352941,
 'sessionid': 1,
 'split': 'train',
 'transcription_text': 'innkalte vararepresentant for buskerud fylke elisabeth skogrand
har tatt sete til behandling foreligger det dagsorden nummer trettifire og sak en er
spørretime muntlig spørretime stortinget mottok mandag meddelelse fra statsministerens
kontor om at utenriksminister jonas gahr støre og statsrådene knut storberget og lars
peter brekk vil møte muntlig spørretime',
 'transcriptionfile': 'data/transcriptions/json/2010-01-06_transcription.json',
 'speakers': [{'speaker_id': 'person.DTA',
   'birth_county': 'Akershus',
   'rep_counties': ['Vestfold'],
   'language': 'nob',
   'dialect': 'east',
   'dob': '1957-05-27',
   'gender': 'M',
   'age': 52}]}
```


Data Fields

- `segment_id (int)`: Unique ID of the segment
- `sessionid (int)`: ID of the session (the specific parliamentary meeting) the segment belongs to
- `meeting_date (str)`: The date of the parliamentary meeting in yyyy-mm-dd format
- `split (str)`: The data split the segment belongs to (`train`, `test` or `eval`)
- `proceedings_text (str)`: The text of the segment as extracted from the parliamentary proceedings. This field is intended for training/fine-tuning weak supervision ASR models. See more on the extraction proceedings of `proceedings_text` below and in the reference article. Note that this string is not necessarily a verbatim rendering of what is said.
- `context_before (str)`: The context immediately preceeding `proceedings_text` in the official proceedings of Stortinget. Intended for prompting as well as adjusting segment boundaries. The number of words of `context_before` varies, but it is within one standard deviation of the mean word length of the `proceedings_text` values (unless the distance to the start of the document is shorter).
- `context_after (str)`: The context immediately following `proceedings_text` in the official proceedings of Stortinget. Intended for adjusting segment boundaries. The number of words of `context_after` varies, but it is within one standard deviation of the mean word length of the `proceedings_text` values (unless the distance to the end of the document is shorter).
- `transcription_text (str)`: ASR transcription of the segment by one of these wav2vec2 models: nb-wav2vec2-300m-nynorsk, nb-wav2vec2-300m-bokmaal. Used in the creation of the SSC to extract segments from the proceedings. It is part of the release to enable adjustments and data cleaning of the extracted segments.

- `score (float)`: The Levenshtein ratio between a normalized version of `proceedings_text` and `transcription_text`. The score was used in the extraction of the `proceedings_text` as described in the reference article. The score can be used as an indication of how verbatim the `proceedings_text` is, provided that the `transcription_text` is relatively accurate. Only segments with a score above 0.5 are kept.
- `duration (float)`: The duration of the segment in seconds
- `num_speakers (int)`: The number of speakers in the segment
- `audio_path (str)`: Relative path to the MP3 file containing the audio of the segment
- `proceedingsfile (str)`: Relative path to the text file with the proceedings text of the session. Included for reproducability
- `transcriptionfile (str)`: Relative path to the JSONL file with the ASR transcriptions of the session. Included for reproducability
- `speakers (list)`: A list of dicts with the speakers in the segment
    - `speaker_id (str)`: The ID of the speaker in the ParlaMint-NO corpus
    - `birth_county (str)`: The birth county of the speaker according to Stortinget's API
    - `rep_counties (list)`: A list of the county or the counties the speaker has represented at Stortinget according to Stortinget's API
    - `language (str)`: The written language (Norwegian Bokmål (`nob`) or Norwegian Nynorsk (`nno`)) used in the proceedings to report the speaker's utterance.
    - `dialect (str)`: The dialect region (`east`, `west`, `south`, `mid` or `north`) of the speaker, inferred from `birth_county` and `rep_counties`.
    - `dob (str)`: The date of birth of the speaker, according to the ParlaMint-NO corpus
    - `gender (str)`: The gender of the speaker (`M` or `F`), according to the ParlaMint-NO corpus
    - `age (int)`: The age of the speaker, calculated from `dob (str)` and `meeting_date (str)`

Data Splits

The data splits of the SSC are based on a different speech corpus with Stortinget data, the Norwegian Parliamentary Speech Corpus (NPSC), which contains 126 hours of audio and manual transcriptions of selected parliamentary meetings from 2017 and 2018. The parliamentary meetings which are part of the eval and test splits in the NPSC are kept aside as eval and test splits also in the SSC. Due to the difference in size between the SCC and the NPSC, the eval and test segments constitute a very small proportion of the total segments (0.28% and 0.26% respectively). It was not entirely clear to the development team of the SSC how to make canonical splits of this dataset, as users may want to use only portions of the dataset. However, they deemed it important to keep aside the eval and test data from the NPSC, as the NPSC contains manual, gold standard transcriptions and may be used to test systems trained on the SSC.

Dataset statistics

- Number of segments: 724 783
- Total duration in hours: 5190
- Number of unique speakers: 729

Distribution of single and multiple speaker segments

Number of speakers	Percentage of segments
1	86.14%
2	12.22%
3	1.63%
4	0.01%

Distribution of written languages

Segments with multiple speakers can contain both written languages. In this table and in the tables for dialect and gender distribution, we have filtered out segments with multiple speakers before calculating the percentages.

Written language	Percentage of segments
nno	11.56%
nob	88.44%

Dialect distribution

Dialect region	Percentage of segments
east	33.94%
mid	9.51%
north	6.05%
south	3.54%
west	23.14%

Note that the dialect region is inferred from the speaker's county of birth and the county they represent.

Gender distribution

Gender	Percentage of segments
F	39.96%
M	59.82%

Dataset Creation

Curation Rationale

The creation of this dataset was motivated by the need for large Norwegian speech datasets for Norwegian ASR, and in particular for weak supervision ASR systems such as Whisper.

Source Data

We refer to the reference article for an in-depth explanation of the data collection and transcription matching procedure, and only a brief overview is given here. See also the transcription matching Github repository.

Data collection

The audio data and transcriptions were collected by the National Library of Norway in the following way: * The audio data was scraped from the Stortinget video archive, and the links to the videos can be found in the document `ssc_v1_0_video_urls.txt`, which is part of this release. * Audio from one of the audio channels was extracted from the videos * The audio segments were transcribed to Norwegian Bokmål and Nynorsk using the following wav2vec2 models: nb-wav2vec2-300m-nynorsk, nb-wav2vec2-300m-bokmaal * Speech segments were identified using Silero VAD and the segments were merged into up to 30 second segments using an algorithm from CLARIN.SI * The text of the official proceedings of Stortinget were taken from a beta release of the ParlaMint-NO corpus * The proceedings text and ASR transcriptions were run through various normalization and data cleaning steps before running a Levenshtein-based matching algorithm, extracting the segment in the proceedings text most similar to the ASR transcription. A score, giving the Levenshtein ratio between the normalized text segments, were also produced. Only segments with a score above 0.5 were kept. The matching algorithm used is borrowed from the ParlaSpeech-HR project with minor modifications. The extracted proceedings transcriptions were kept in their original form, without normalization and cleaning.

The speaker metadata were collected from the ParlaMint-NO XML files and the Stortinget API by Phoebe Parsons at the Norwegian University of Science and Technology. The county of birth and county the speaker represents were mapped to dialect regions with a mapping created by Parsons. The code for the metadata extraction is available in this repository.

Who are the source language producers?

The audio data are recordings of speakers at Stortinget (mostly MPs and ministers). The `proceedings_text` transcriptions are extracted from the official proceedings of Stortinget, which are written by the Stortinget stenographers. The `transcription_text` strings are ASR transcriptions.

Personal and Sensitive Information

The audio data, proceedings and metadata, including metadata about the individual speakers, are freely distributed by Stortinget and are public domain under the Norwegian Licence for Open Government Data (NLOD) 2.0, and is not assumed to contain any information exempt from public disclosure.

Considerations for Using the Data

Social Impact of Dataset

This dataset increases significantly the amount of openly available, transcribed speech data in Norwegian and can be used to improve the accuracy and generalization of Norwegian speech recognition systems and other kinds of speech technology. Norwegian is a relatively low-resource language in the speech domain. If ASR systems can generalize to more Norwegian dialects and voices, they can be used effectively by more people.

Additional Information

Dataset Curators

- Per Erik Solberg (National Library of Norway)
- Pierre Beauguitte (National Library of Norway)
- Phoebe Parsons (Norwegian University of Science and Technology)
- Per Kummervold (National Library of Norway)
- Freddy Wetjen (National Library of Norway)

The creation of the SSC is in part funded by the SCRIBE project

Licensing Information

CC0

Citation Information

```
@inproceedings{solberg2023large,
  title={A Large Norwegian Dataset for Weak Supervision ASR},
  author={Solberg, Per Erik and Beauguitte, Pierre and Kummervold, Per Egil and Wetjen,
Freddy},
  booktitle={Proceedings of the Second Workshop on Resources and Representations for
Under-Resourced Languages and Domains (RESOURCEFUL-2023)},
  pages={48--52},
  year={2023}
}
```