Nasjonalbiblioteket

Norwegian Parliamentary Speech Corpus

The Norwegian Language Bank

1. About the corpus

General introduction

The Norwegian Parliamentary Speech Corpus (NPSC) is a speech corpus made by the Norwegian Language Bank at the National Library of Norway in 2019-2021. The NPSC consists of recordings of speech from Stortinget, the Norwegian parliament, and corresponding orthographic transcriptions to Norwegian Bokmål and Norwegian Nynorsk, as well as various metadata about the speakers. All transcriptions are done manually by trained linguists or philologists, and the manual transcriptions are subsequently proofread to ensure consistency and accuracy.

This corpus is primarily intended as an open-source dataset for ASR development. Recordings from Stortinget are well suited for an open-source resource of this kind. Firstly, the audio from Stortinget is shared with an open license, and consequently, there are no copyright restrictions when reusing and resharing the audio. Secondly, all speakers are public figures and information about their age, place of birth etc. is, in most cases, available from public sources. Thirdly, dialects are widely used in Stortinget, and speakers come from all regions of the country, which means that we get a good dialect distribution in the material. Fourthly, all plenary meetings at Stortinget have official proceedings. While these sometimes deviate a bit from the spoken words in order to function well as a written text, they still render the speech quite faithfully. These proceedings are a handy tool for the transcribers, and they may also be beneficial to users of the corpus. The proceedings are therefore shared together with the transcriptions and the audio.

The audio files in the corpus contain the speech of entire days of plenary meetings from 2017 and 2018, or, if a meeting lasts more than six hours, the first six hours of a day. Since the entire audio files become quite large, we also provide individual audio files for each sentence.

License

The NPSC comes with a CC0 license, i.e., it is public domain and can be used for any purpose and reshared without permission.

Nasjonalbiblioteket

Versions and technical specifications

This is version 1.1 of the NPSC

Version 1.0

There were beta releases of parts of the NPSC in 2020 and the spring of 2021. Be aware that the formatting of the transcription and metadata files was different in the beta releases from the current version. Also, we have corrected errors in the transcriptions and preprocessed in multiple ways since the last beta release. We therefore recommend that users replace transcription files from previous versions with this version. The audio files from the previous versions are not altered in any way, so they may be kept.

The following audio files with corresponding transcriptions and metadata, are included in the NPSC. The second column gives the duration of the audio files in an *h:mm:ss* format.

Audio file	Duration
20170208-095517.wav	2:29:09
20170209-095509.wav	1:40:17
20170216-095707.wav	1:48:14
20170314-095510.wav	3:05:47
20170405-095525.wav	1:58:46
20170613-085456.wav	6:10:01
20170207-095506.wav	1:14:02
20171007-125517.wav	1:19:36
20171211-095523.wav	6:10:10
20171219-095511.wav	6:10:11
20170215-095508.wav	3:44:33
20171018-095533.wav	2:45:57
20180201-095509.wav	1:40:45
20170510-095755.wav	4:00:16
20170615-085510.wav	6:10:10
20170419-095601.wav	1:38:47
20170503-095709.wav	3:12:26
20170403-115516.wav	3:37:39
20171213-095535.wav	5:30:09
20170222-095521.wav	4:53:11
20170426-095547.wav	3:06:10
20171122-095628.wav	3:00:51
20170322-095517.wav	3:40:36
20180321-095533.wav	3:12:56
20180410-095617.wav	6:10:10
20180316-085409.wav	1:36:08

20180404-095533.wav	1:43:58
20171012-095507.wav	1:53:04
20171208-085509.wav	2:38:48
20170516-095522.wav	1:32:15
20180307-095547.wav	3:23:11
20180411-095502.wav	3:28:04
20170110-095504.wav	4:50:33
20171024-095522.wav	0:10:28
20170323-095515.wav	3:20:47
20180611-095427.wav	6:10:11
20180109-095530.wav	1:47:01
20180615-085438.wav	6:10:10
20180530-095508.wav	2:40:29
20180601-085543.wav	4:14:07
20180613-095502.wav	6:10:11
Total	**140:20:16**

Table 1: File names and durations

Note that the total duration of the sentence-segmented files is somewhat shorter than the total in the table above (approximately 126 hours), as pauses are not included.

For audio files of meetings lasting more than six hours, the audio is cut at approximately 6 hours and 10 minutes, usually mid-sentence. So the last sentence of the longest audio files is usually not complete.

The files containing transcriptions and metadata are formatted in JavaScript Object Notation (JSON). The parliamentary proceedings are regular text files. Both are encoded in UTF-8. The audio files are wav files with the following technical specifications:

Format	PCM
Format settings	Little / Signed
Bit rate mode	Constant
Bit rate	1 536 kb/s
Channels	2[1]
Sampling rate	48 kHz
Bit depth	16 bits

Table 2: Audio file specifications

[1] The parliament lectern has two parallel microphones that seem to be about 15 cm apart. We assume that the two channels correspond to each of these microphones.

Unfortunately, the specifications above do not reflect the actual quality of the recordings, as the audio in the videos from Stortinget, from which the audio files are extracted, is somewhat compressed.

Version 1.1

The following changes were made in this version of the NPSC:

- The data was split in an official training, evaluation and test set
- Manual dialect annotations were added for each speaker
- The end time of one sentence in *20171208* (sentence_id 45886), was changed, as a 30 minute break was included in the sentence time span in version 1.0. The corresponding audio file (20171208-085509_6122400_6124160.wav) was shortened accordingly.
- This document was updated to reflect these changes
- Some of the metadata in the transcriptions of 20171213 were lacking in the json transcription files. These are added in this version.

Transcription pipeline

Figure 1 is a visual representation of the transcription pipeline of the NPSC project:

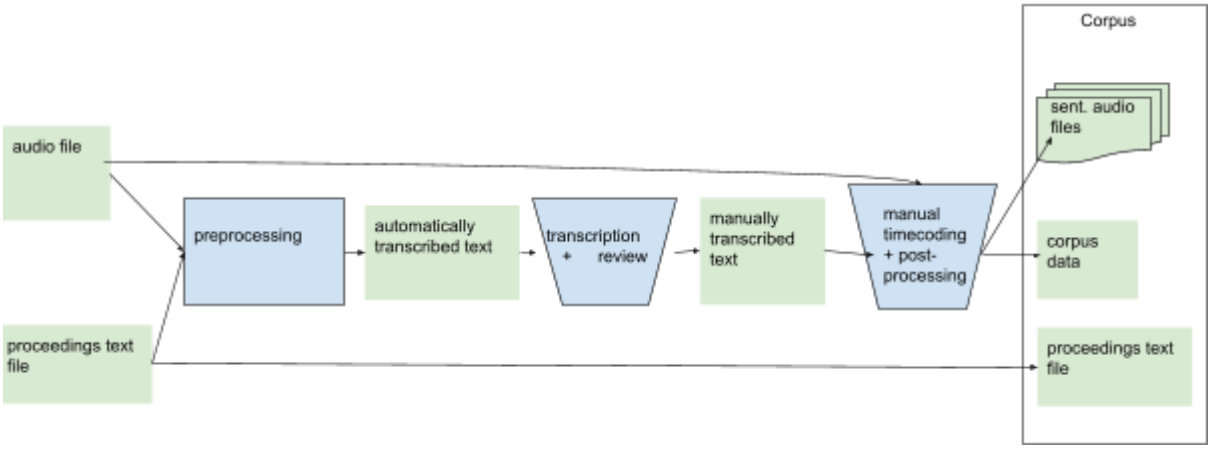

Figure 1: Transcription pipeline

Preprocessing

An audio file of a day of parliamentary speech is run through Google Cloud Speech-to-Text. This produces an automatic transcription to Norwegian Bokmål. A script subsequently runs through the automatic transcription and the proceedings text file to check for sequences of words with a high similarity (using Levenshtein distance). We assume that if there are sequences of words in the ASR output and the proceedings with a high similarity, but not identity, the proceedings text is the correct version. In such cases, the script swaps the words in the ASR output with the corresponding words in the proceedings file. This produces a somewhat improved automatic transcription. The proceedings file renders the speech of

certain speakers in Norwegian Nynorsk (more on this below). In such cases, the script will often substitute some of the Bokmål words from the ASR output with Nynorsk words.

Transcription + review

The automatic transcription is presented to a transcriber in a web-based GUI which allows the transcriber to listen to the audio and read and correct the transcription. The transcriber corrects the automatic transcription and segments it into sentences. After this correction, the timestamps from the ASR will no longer be correct. This will be fixed at a later stage. Each sentence is annotated with the name of the speaker.

MPs who have Bokmål as their written language of choice will be cited in Bokmål in the proceedings, and Nynorsk MPs will be cited in Nynorsk. We follow this practice, as it often makes sense based on their dialect and it makes it possible to benefit from the fact that the preprocessing step partly produces Nynorsk when there is Nynorsk in the proceedings.

When MPs use dialect words or forms that do not correspond to the written standards of Bokmål or Nynorsk, the transcribers follow certain conventions to ensure consistency and that the transcriptions are as faithful as possible to what is actually said without breaking with the norm. In certain cases, the transcribers annotate a word as being nonstandard and give an alternative, standard form. More on this below.

When the transcription is completed, another staff member at the Language Bank reads through the transcription, listens through the audio and corrects the transcription when needed.

Manual timecoding + postprocessing

Once the transcription is through review, the start and end time of each sentence is checked and corrected manually using the software Elan.

The audio and the transcription are then run through a number of postprocessing steps: Audio files for each sentence are produced, based on the manually corrected time codes, using Pydub.

The annotation of the speaker is extracted and matched with names in a list of MPs and members of government. This list contains date of birth, place of birth, URIs and other metadata extracted from Wikidata. This information is retrieved and added to the corpus data JSON files, together with the transcription, time codes, annotations of non-standard language and other relevant metadata described further in the next section.

Corrections, normalization and machine translation

At the end of the project period, the transcriptions were run through a number of additional postprocessing steps: known misspelled words in the transcriptions were corrected and the

transcriptions were run through a normalization grammar in order to produce an alternative, normalized version of the transcription. Furthermore, we have added a version of each sentence which is machine translated to the other written standard (i.e. Bokmål for Nynorsk transcriptions and Nynorsk for Bokmål transcriptions). The normalization is further described here and the translation is described here. The Python scripts used in this final postprocessing pipeline can be found in this release in *project_files/postprocessing_scripts*.

2. Overview of the transcription conventions

Guiding principles

The following principles have guided our transcription work:

1. **Consistency**. Similar linguistic phenomena should be treated similarly across transcriptions regardless of who performed the transcription
2. **Standardized orthography.** The transcriptions should follow standardized orthography whenever possible.
3. **Faithful rendering of speech.** The transcriptions should render the pronunciation as faithfully as the orthographic standard allows. When the norm allows for multiple transcriptions of a word, we always choose the one which more closely reflects what the speaker said.
4. **Flagging of non-standard speech.** If the speech deviates significantly from the written standard, either due to dialects or for other reasons, this should be flagged. If a non-standard form of a word is given in the transcription, the transcriber should also provide a standardized semantic equivalent.

A number of measures have been taken to ensure consistency (1): Firstly, the transcribers have followed detailed transcription guidelines, of which this is only a summary. See *project_files/NPSC_transcription_guidelines.pdf* for the full version of the guidelines (in Norwegian). These guidelines are written by some of the transcribers during the course of the transcription. Secondly, transcribers discuss issues as they come up, either in in-person meetings or via chat. Thirdly, all transcriptions have been reviewed in their entirety by a colleague. Finally, the transcribers check and maintain word lists whenever they need to write a non-standard form and provide a semantic equivalent (4), to ensure as much as possible that transcribers treat similar cases similarly.

Since these transcriptions are orthographic, users need to know that they actually conform to the written norm of Bokmål and Nynorsk (2). By strictly adhering to the official standard, we also ensure consistency (1). The transcriptions should nevertheless be as close as possible to the actual pronunciation (3), since this is important, e.g., for training acoustic models. The Bokmål standard, and to a somewhat lesser degree the Nynorsk standard, allow for some variation. For example, many nouns can be either masculine or feminine, depending on the dialect, and both forms are usually allowed in Bokmål. However, given the differences of the

dialects in Norway and the widespread use of them in Parliament, you sometimes have to make a choice between using standardized orthography (2) or rendering the speech faithfully (3). In such cases, we have chosen different practices for lexical words (nouns, regular verbs, adjectives etc.) and function words (auxiliaries, articles, prepositions etc.): In the case of lexical words, we write a non-standard form close to the actual pronunciation, flag the word as having a non-standard spelling (4), and provide a semantic equivalent which conforms with the standard. We deemed this practice not feasible for function words, however, as they are so frequent and are realized so differently across dialects and individuals. Function words are therefore always written according to the standard, but in cases where the standard and the actual pronunciation diverge widely, the function word is flagged as having a *phon_ort_discrepancy*. This flagging is not done as consistently as the flagging of lexical words, as it is difficult to formulate exact criteria for what is sufficiently divergent. We return to both these cases in some more detail below.

General transcription conventions

Sentence-segmentation

Each full sentence is a segment in the transcription, and the start and end timecodes of each sentence are manually adjusted. Some non-sentence units may have a syntactic function similar to a sentence, e.g. *ja*, 'yes', and when a speaker is introduced by their name. Units of this kind are also treated as sentences in the transcription. There are cases where there are several possible ways of segmenting a transcription into sentences. The transcribers have guidelines for what to do in such cases, in order to ensure as much consistency as possible (principle 1).

Non-standard language

Non-standard language is defined in the NPSC project as words which are not part of the official norm of Bokmål or Nynorsk (depending on the written norm used in the transcribed utterance), or whose pronunciation deviates significantly from the standard orthography.[2]

Two points in this definition need some further explanation: Firsty, the definition only covers non-standard *words*; it does not cover syntax. Therefore, if a speaker uses syntactic constructions which may be considered non-standard, but uses only standard vocabulary, this will not be marked or altered in any way. For example, while Bokmål has a distinction between nominative *de* and accusative (or rather, *oblique*) *dem*, 'they', a good number of dialects use *dem* in all syntactic contexts, including in subject position. Having *dem* as a subject would be considered non-standard in writing, however. Since *dem* is part of the Bokmål vocabulary, we allow *dem* in subject position in our transcription. The reason for not taking non-standard syntax into account is that it would make it more difficult to render the

[2] To decide whether a word is part of the standard or not, we use https://ordbok.uib.no/ and https://naob.no/ as well as various word lists on the website of the Language Council.

pronunciation as faithfully as possible (principle 3). Also, producing well-written prose is outside the scope of this project.

Secondly, the pronunciation of a word needs to deviate *significantly* from the standard in order to be considered non-standard in the transcriptions. It is of course difficult to define precisely what is considered a significant difference. Nevertheless, we deemed it necessary to have a requirement like this. Most Norwegian dialects deviate from the orthography to some degree, and we needed to avoid marking a significant proportion as non-standard. Dialectal inflectional variants are usually not considered non-standard. For example, despite the dialectal ending of *gutad'n*, this word would be transcribed as *guttene*, 'the boys', in Bokmål or *gutane* in Nynorsk without any special marking.

When marking non-standard language, we distinguish between function words and lexical words, as mentioned previously. For non-standard pronunciations of function words, we write a standard form of the word, but mark the token as having a *phon_ort_discrepancy* (see this section for details on how this is represented). Words in the following classes are considered function words: pronouns, determiners (except numerals), prepositions, complementizers, conjunctions, auxiliary verbs, interrogative adverbs, negation and some temporal and locative expressions meaning 'then', 'now', 'here' and 'there'.

Non-standard lexical words, i.e. all non-standard words apart from the function words, are treated differently. The word is transcribed with a form close to the pronunciation in *sentence_text* and *token_text*, but the token is flagged as having a *non-standard_spelling.* Furthermore, a standardized equivalent word is provided in the token field *standardized_form* (cf. the description of the token-level transcription files). The transcribers maintain a spreadsheet with all non-standard forms and their equivalents for the sake of consistency (principle 1). The non-standard form is not written in a phonetic script. We try to follow the general principles of Bokmål or Nynorsk orthography, to the extent that it is possible, when writing the non-standard form. Often, a word is non-standard when transcribing Nynorsk, but the same word would be standard in Bokmål (or, less often, vice versa). In such cases, we usually choose the Bokmål word as the non-standard form and give the Nynorsk equivalent as the standardized form (or vice versa).

It is sometimes impossible to find a one-word equivalent to a non-standard word. When this is the case, the standardized form will be an MWE with space replaced by underscore. For example, *opprettholde,* 'sustain', which is not allowed according to the Nynorsk norm, gets the standardized equivalent *halde_ved_lag*. In some cases, we have MWEs as the non-standard form. This happens when a word in the MWE is not used outside of MWEs and it is therefore not possible to find a standardized equivalent for the word in question. *Tross* in *på tross av*, 'despite', is non-standard in Nynorsk, but it is also semantically void in itself and it is therefore difficult to find a good equivalent. In such cases the whole MWE forms one token, with the form *på_tross_av* and the standardized form *trass*. In the sentence-segmented version of the corpus, the underscores are replaced by space, however. We have tried to limit this latter case to a minimum due to the exceptional tokenization.

Note, finally, that Nynorsk transcriptions contain non-standard material to a much higher degree than the Bokmål transcriptions. An important reason for this is that the Nynorsk standard disallows vocabulary of Low German origin, at least to a large extent. This ban does not correspond to actual usage of words in the spoken language: Words of Low German origin are used frequently by most people in their spoken language, including people from Nynorsk areas. A large proportion of the non-standard words in the Nynorsk transcriptions are of Low German origin.

Hesitations, interruptions and inaudible speech

Hesitations are explicitly transcribed. Vocalic hesitations are transcribed as <*ee*>, nasal hesitations as <*mm*> and other non-linguistic speech sounds such as cough are transcribed as <*qq*>. These are also marked with *special_status: HESITATION*.

Interrupted speech is also explicitly transcribed. The transcription contains a letter representation of the string produced, which often, but not always, corresponds to a partial word. The token is marked with *special_status: INTERRUPTED*.

Finally, parts of the recordings with inaudible or overlapping speech are transcribed as *<INAUDIBLE>.* These are marked with *special_status: INAUDIBLE* or *special_status: OVERLAPPING*.

For more information on how these features are formally represented in the transcription file, see the description of the token-level transcription files.

3. Speaker annotations

When transcribing speech, the transcribers annotate each sentence with the name of the speaker. These names have later been checked against a list of MPs and members of government derived from Wikidata using their SPARQL endpoint. For all names found in that list, we have derived from Wikidata information about their gender, date of birth, place of birth, and the county and region of the place of birth, as well as their electoral district, if this information is present there. This information may be used to predict which dialect group they most likely belong to. Wikidata URIs are given for the speaker and for their place of birth, in case users of the corpus want to retrieve more metadata.

Dialect annotations

In version 1.1, dialect annotations were added to each speaker. An annotator listened to the sentence with the longest duration of each speaker in the NPSC, and decided which region the speaker was from (Eastern Norway, Southern Norway, Western Norway, Trøndelag, Northern Norway). Not all dialects fit neatly into this regional division. In particular, dialects in

Trøndelag and the northern parts of Møre og Romsdal (a county in Western Norway) are very similar. Dialects from this area are classified together with Trøndelag dialects. Similar cases could arise in other parts of the country, but the annotator is only aware of other cases in the NPSC. In a few cases, speakers have dialect features of different dialects. In such cases, the annotator has opted for the dialect which seemed most dominant to them.

4. Normalization

The manual transcriptions were so-called “spoken domain” transcriptions, which means that the transcribers represented the speech as faithfully as the orthography allows. In particular:

- Numbers, years and dates are written with letters as they are pronounced: *hundre og femti tusen*, ‘hundred and fifty thousand’, *første juli tjueatten*, ‘july first twenty eighteen’.
- Abbreviations are not used (unless the speaker actually pronounces the abbreviation)

For some use cases, it may be useful to also have “written domain” transcriptions, where numbers, and years are written with digits, dates are written using a standardized format and common abbreviations are used. As explained above, we have written normalization grammars that produce a normalized, “written domain” version of the transcriptions. Note that this is a completely automatic procedure, and may therefore contain errors. We have, however, done quality spot checks and subsequent refinements to fix errors we have identified.

This is a summary of the normalizations:

- Numbers are written as digits: *hundre og femti tusen -> 150000*
- Years are written with digits: *tjueatten -> 2018*
- Dates are written with digits using the format *<day>.<month>.<year>: første juli tjueatten -> 1.7.2018*
- Real numbers are written with a comma (which is standard in Norway): *2,5*
- Standard abbreviations are used. See the list of abbreviations in the grammar source code in *project_files/postprocessing_scripts/grammars/abbrev_grammar.py*.
- Percentages are written as follows: *to prosent -> 2%*. Note that the conventional way of writing percentages in Norwegian is with a whitespace between the number and the percent sign. Because of tokenization issues, we chose to drop the whitespace between the number and the percentage sign.

The normalized version of the corpus (as well as the non-normalized version) exists in a word-tokenized and a sentence-tokenized version, as explained below.

As described in the overview of the transcription guidelines, non-standard lexical words are annotated with a standardized semantic equivalent. To maximize the number of cases that the normalization scripts apply to, they use the standardized equivalents whenever present. In particular, the number normalization grammar converts number expressions which contain

non-standard number words. There are two paradigms in use for numbers between 21 and 99 in Norwegian, both of which are in use orally, but only one is standardized in writing: *fireogførti* and *førtifire* both mean 44, but only the latter is allowed in written Norwegian. Since the normalization grammars use the semantic equivalents, the normalization applies to numbers from both paradigms.

5. Machine translation

Since we follow the choice of written standard of the speaker and since the majority of Norwegians use Bokmål, only about 12% of the corpus is in Nynorsk. To increase the amount of Nynorsk in the corpus, we have machine translated the Bokmål transcriptions to Nynorsk. Since we already had a procedure for machine translation in place, we also translated the Nynorsk sentences into Bokmål. These translations are not in any way a substitute for manual transcriptions, as they are produced automatically without regard to the audio. However, it was a relatively easy material to make, so we chose to add it. It should be considered an experimental material.

We used the rule-based translation system Apertium. In this system you can specify if infinitives in Nynorsk should end in *-e* or *-a* (both of which are allowed in Nynorsk). We chose *-e*. When Apertium encounters a word it does not recognize, it marks it with a Kleene star, e.g. **ECT*. We left this marking untouched. Note that metatags also get this marking, e.g. <**ee*> and <**INAUDIBLE*>.

The input to the translation is the normalized ("written domain") version of the sentences. Before the machine translation, all non-standard words are substituted with their standard semantic equivalent to minimize the amount of non-recognized words. The translations are sentence-tokenized, but not word-tokenized.

The corpus contains a few English sentences. These are not translated.

The source code for the translations can be found in *project_files/postprocessing_scripts/oversettelse.py*.

6. Description of files and directories

Naming conventions and directory structure

All files in the corpus are uniquely named, so it is not necessary to keep the directory structure described here.

As explained above, the starting point of a transcription is the recordings of a day of plenary meetings in parliament, or the first six hours and ten minutes of the day in cases where the

plenary meetings extend beyond six hours and ten minutes. A directory is created for each transcribed day in parliament, and the directory name is the date, e.g. *20170207/*. In addition to the dated directories, there is also a directory called *project_files/*, containing the transcription guidelines, the *NPSC_speaker_data.json* file, with speaker metadata, as well as an Sqlite-version of the transcriptions and the postprocessing scripts used in the final phase of the project. We will first present the content of the dated directories and then *project_files/*.

A dated directory contains five files, named with the date and the start time of the recording and a file signature: *yyyymmdd-hhmmss.**, e.g. *20170207-095506.ref*, or only the date, e.g. *20170207_sentence_data.json*, as well as a directory called *audio/*:

- *20170207-095506.ref* is a text file with the official proceedings from Stortinget. Note that when the meetings last more than six hours and ten minutes, the audio and the transcription only cover the first six hours and ten minutes, while the proceedings file covers the entire meeting.
- *20170207-095506.wav* is the audio file of the entire meeting or the first six hours and ten minutes. Note that files containing only the first part of a long meeting are cut at six hours and ten minutes, which is often in the middle of a sentence. The last sentence of the audio file and the transcription will therefore be incomplete in these cases.
- *20170207_sentence_data.json* sentence-tokenized transcriptions in non-normalized form and normalized form, as well as the machine-translated version of the sentence.
- *20170207_token_data.json* contains a word-tokenized version of the transcriptions, including metadata about the word. In particular, information about non-standard spelling and semantic equivalents is found here.
- *20170207_normalized_token_data.json* contains the normalized version of the word tokens.
- *audio/* contains individual wav files with the audio of each sentence. They are also named according to the date and start time, but they also have the start and end time in milliseconds of the sentence they contain: *yyyymmdd-hhmmss_<starttime>_<endtime>.wav*, e.g. *20170207-095506_302650_306000.wav*.

Data splits

In version 1.1 we have split the data into an official training, evaluation and test set. Since some users may be interested in the context beyond the sentence, we chose to split the data on meeting days rather than on sentences. We tried to stay as close as possible to a 80-10-10 percent split, calculated from the total duration of speech, where pauses are excluded. The information about splits is recorded in the *sentence_data* and *token_data* files (see below). These are the dates included in the test and evaluation sets:

Test	20171219, 20180530, 20171122, 20170207
Evaluation	20180611, 20180201, 20170209, 20180307, 20180109

Table 3: Data splits

To ensure that the splits are as balanced as possible, we identified three features that should be as similar as possible across splits: the percentage of Nynorsk, the percentage of female speakers, and the average word length per sentence. We have not taken the dialect distribution into account when deciding on the splits. However, we have verified that all the dialect areas are represented in reasonable proportions in each split. The table below lists the duration, the percentage of Nynorsk and female speakers, the average sentence length and the percentage of sentences from each dialect area.

	duration	Nynorsk	female	sentence length	Western N.	Eastern N.	Southern N.	Northern N.	Trøndelag
test	12.3	13	39.9	18	35	44.4	4.6	9.4	6.5
eval	13.1	12.7	41.8	18	32.2	43.6	7.6	7.7	8.8
train	100.3	12.8	37.7	18.7	26.5	46.1	6.5	11.9	9.1
total	125.7	12.8	38.3	18.6	27.9	45.6	6.4	11.2	8.8

Table 4: Split statistics

The content of the *sentence_data* file

The *sentence_data* json file contains sentence-segmented transcriptions in non-normalized and normalized form, machine-translated sentences and relevant metadata. On the top level, the sentence_data file is a dictionary with the following keys and values:

- *meeting_date*: Date of the meeting in a *yyyymmdd* format
- *full_audio_file*: The name of the audio file containing the audio from the entire meeting
- *proceedings_file:* The name of the file with the official proceedings produced by Stortinget
- *duration:* The duration of the full audio file in milliseconds
- *transcriber_id*: The id of the transcriber of the current transcription
- *reviewer_id:* The id of the reviewer of the current transcription
- *data_split:* Which data split (*test*, *eval* or *train*) the meeting belongs to
- *sentences:* A list containing the sentence-tokenized transcriptions with metadata, formatted as dictionaries.

The sentence-dictionaries contain the following keys and values:

- *speaker_name:* The full name (*first_name last_name*) of the speaker of the sentence
- *speaker_id:* The id of the speaker of the sentence. More metadata about the speaker can be found in the file *NPSC_speaker_data.json*, described below

- *sentence_id*: The id of the current sentence. This id is referred to in the token metadata in the word-tokenized versions of the corpus
- *sentence_language_code:* The language code of the sentence. This is mostly either *nb-NO* (Bokmål) or *nn-NO* (Nynorsk), but there are occasional instances of English, in which case *en-US* is used. We transcribe the sentences of a speaker using the same written standard as the official proceedings use for the speaker (which is the speaker's own chosen standard). However, transcribers can mark sentences with a different language code, i.e. in the case of quotes. In that case, the *sentence_language_code* will follow the explicit marking, not the speaker's language. Transcribers can also mark some, but not all tokens in a sentence with a different language code. This will affect the language code of the marked tokens in the word-tokenized version of the sentence (see below), but not *sentence_language_code*.
- *sentence_text:* The transcribed text string of the sentence in non-normalized form. This is the text of the manual transcriptions, without any postprocessing (apart from corrections of known errors). It may contain interrupted words, non-standard words and function words with a pronunciation deviating from the written form. Information about this is found in the word-tokenized version of the sentence.
- *sentence_order:* This field contains a number indicating the order of the sentences in the meeting
- *audio_file:* The name of the sentence-segmented audio file in *audio/* containing the audio of the current sentence.
- *start_time*: The start time of the sentence in milliseconds
- *end_time:* The end time of the sentence in milliseconds
- *normsentence_text:* The sentence in normalized form (see above for an explanation of the normalization procedure)
- *transsentence_text:* The sentence machine-translated to the Bokmål if *sentence_language_code* is *nn-NO*, or Nynorsk, if *language_code* is *nb-NO* (see above for an explanation of the translation procedure)
- *translated*: Indicates whether a machine-translated version has been produced or not. If *sentence_language_code* is *nb-NO* or *nn-NO*, a machine-translated version is produced, *translated* has the value *1*. If *sentence_language_code* is *en-US*, the value is *0*. In that case, *transsentence_text* will be a copy of *sentence_text*.

The content of the token_data file

The token_data json file contains word-tokenized transcriptions in non-normalized form. The lists of tokens are grouped by sentence. The file is structured in a similar fashion to the sentence_data file: The top level is a dictionary with the same keys as the top level of the sentence_data dictionary. The values of the metadata keys are the same, but *sentences* contains a list of dictionaries which differs from the sentence_data file. These are the key-value pairs of the dictionaries in *sentences*.

- *speaker_name*: see this key in the description of the sentence_data file
- *speaker_id:* see this key in the description of the sentence_data file
- *sentence_id:* see this key in the description of the sentence_data file

- *sentence_order*: see this key in the description of the sentence_data file
- *audio_file:* see this key in the description of the sentence_data file
- *sentence_language_code*: see this key in the description of the sentence_data file
- *tokens*: A list of dictionaries containing the tokens of the sentence.

The *tokens* dictionaries contain the following key-value pairs:

- *token_id*: A unique identifier for the token
- *token_order:* The order of the token in the sentence
- *token_text*: The text string of the token
- *nonstandard_spelling*: *0*, if *token_text* is spelled according to the official norm of either Bokmål or Nynorsk, depending on the token *language_code*; *1* if the word does not follow the norm. In the latter case, there is a standardized, semantic equivalent form in *standardized_form*
- *standardized_form*: If *nonstandard_spelling* is *1*, the value is a string with a standardized, semantic equivalent of the nonstandard word given in *token_text*. If not, the value is *null*
- *special_status:* This field is used for tokens that do not represent a complete, audible word. It can have the string values *INAUDIBLE* or *OVERLAPPING*, in which case *token_text* contains the string <*INAUDIBLE*>. In the case of hesitations, it has the string value *HESITATION*. Incomplete or interrupted words will have the string value *INTERRUPTED*. In all other cases, the value is *null*.
- *language_code*: The value is equal to *sentence_language_code*, unless the token is part of a span marked by the transcriber with a different language code, in which case the explicit annotation is the token's *language_code*.
- *phon_ort_discrepancy*: We have decided not to treat dialectal pronunciations of function function words (prepositions, articles, pronouns and auxiliary verbs) in the same way as other non-standard words (i.e. by providing a standardized form), as they are very frequent and vary to a large degree. Instead, the transcribers have transcribed function words with a normalized form in all cases. However, the transcribers have flagged function words whenever the phonology-orthography discrepancy is large. Words flagged in this manner will get the value *1* in this field. All other words have the value *0*.
- *sentence_id*: the *sentence_id* of the sentence the token belongs to. This id is always equivalent to the sentence-level *sentence_id*.

The content of the normalized_token_data file

The *normalized_token_data* file is organized in a similar fashion to the *token_data* file, but contains the tokens of the transcriptions after the normalization grammars have been applied to them. This implies that some of the tokens in the *tokens* lists are different from the *token_data* file. Also the token-level dictionaries differ from the token-level dictionaries in *token_data* in two ways:

1. The key *normtok_id* replaces *token_id*. This key contains the unique identifier of the token. When the token is unaffected by the normalization, i.e. when *converted* has the value *0* (see immediately below), the *normtok_id* is equal to the *token_id* in *token_data*. If the token is converted by the normalization grammar, the id has the

format *s<number1>e<number2>*, e.g. *s549381e549383*, where *<number1>* is the *token_id* of the start token of the converted span in the non-normalized token list, and *<number2>* is the end token of the span. In the case of one-to-one conversion, *<number1>* and *<number2>* will be the same number, but when the conversion is many-to-one, *<number1>* will be lower than *<number2>*. (The normalization scripts do not support one-to-many conversions.)

2. There is a key *converted*. Tokens which are unaffected by the conversion, have the value *0*, while tokens affected by the conversion have the value *1*.

The *project_files/* directory

The *project_files* directory contains files pertaining to the project as a whole, in particular the transcription guidelines (in Norwegian only), the speaker annotation file *NPSC_speaker_data.json*, *stortinget_speech_corpus.db*, an Sqlite3 version of the transcriptions and metadata, and *postprocessing_scripts/*, a directory with code used in the final postprocessing of the corpus. These files and this directory will be described in turn below.

The transcription guidelines

The transcription guidelines, *NPSC_transcription_guidelines.pdf*, is a document written by and for the transcribers. Note that it is intended as a manual for transcription, not as documentation for the final product, so the formatting of cases of non-standard speech etc. does not correspond to the formatting in the corpus files described here, since the actual transcription files have been converted. However, the linguistic choices described in this document are implemented in the corpus. Note that this document is written in Norwegian Bokmål. Links to internal documents have been removed.

The *NPSC_speaker_data* file

NPSC_speaker_data.json contains metadata for all the speakers in the corpus. The top level is a list. This list contains one dictionary for each speaker. The speaker dictionaries have the following key-value pairs:

- *speaker_id*: The identifier of the speaker. This is the same id as is found in the *speaker_id* fields in the various transcription data files described above.
- *speaker_name*: The name of the speaker written as *<first name> <opt. additional first/middle names> <last name>*, e.g. *Torbjørn Røe Isaksen*.
- *speaker_URI*: URI of the entry of the speaker in Wikidata
- *date_of_birth*: The date of birth of the speaker, extracted from Wikidata, on a *yyyy-mm-dd* format
- *place_of_birth*: Name of the birthplace, if found in Wikidata, or *null*
- *pob_URI*: The Wikidata URI of the place of birth, if available. Or *null*
- *pob_county*: The county of the place of birth, as extracted from Wikidata. When the place of birth is unknown, the value is *null*

- *pob_region*: The region of the place of birth, as extracted from Wikidata. When the place of birth is unknown, the value is *null*
- *electoral_district*: The electoral district of the speaker, as extracted from Wikidata. When the electoral district is unknown, the value is *null*
- *ed_URI:* The Wikidata URI of the electoral district. When the electoral district is unknown, the value is *null*
- *gender:* male or female
- *chosen_language:* The chosen written standard (*nb-NO* or *nn-NO*) of the speaker, i.e. the written standard used to render their speech in the official Stortinget transcripts.
- *dialect*: The region the dialect of the speaker belongs to, according to a human annotator. Note that the regional division is the same as for *pob_region*.

One speaker dictionary has different values from the other speaker dictionaries, namely *speaker_id* 21 with *speaker_name unknown*. Sentences for which the transcriber does not know who is speaking, are annotated with this speaker. In this speaker dictionary, all values are *null*, with the exception of *speaker_id*, *speaker_name*, *chosen_language* (which is *nb_NO*), and *dialect* (which is *unknown)*.

The *stortinget_speech_corpus* database file

stortinget_speech_corpus.db is the Sqlite database from which all the information in the various transcription files and the speaker annotation file are extracted. This file can be used as an alternative way of accessing the transcription data. The architecture is given in figure 2. The information about the different columns can be deduced from descriptions of the content of the various files above.

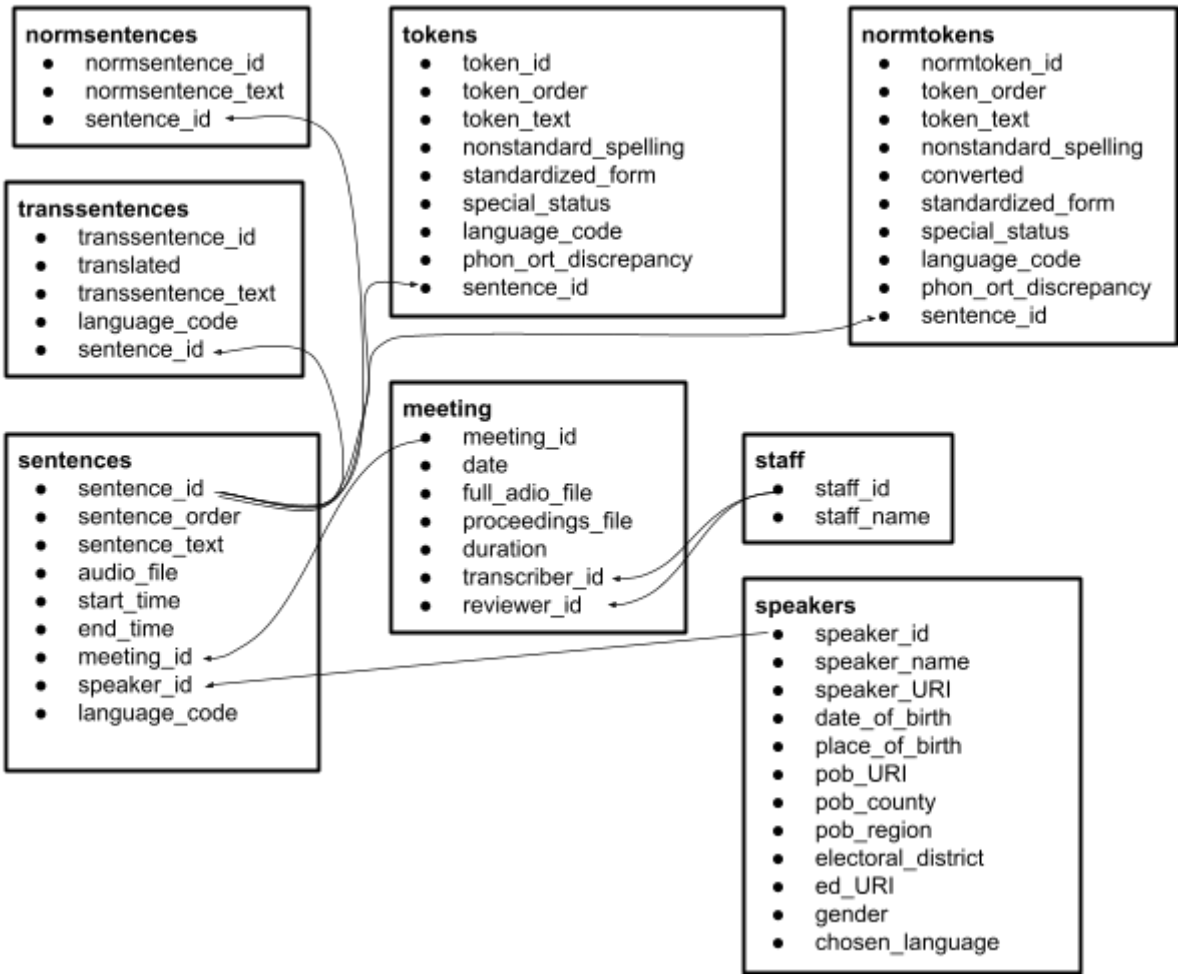

Figure 2: Architecture of the Sqlite database

Note that the *staff_name* column in the *staff* table has been anonymized.

In version 1.1, the *speakers* table has a *dialect* column in addition to the columns in *table 2*, and the *meeting* table has a *data_split* column.

The content of *postprocessing_scripts/*

The *postprocessing_scripts/* folder contains the scripts for corrections of errors, normalization and machine translation, which were run on the corpus after all the transcriptions were added. *pipeline.py* imports the relevant modules and runs them in sequence. These scripts are not fully documented, and it is not our intention that users run them again. We share them to be fully transparent about these postprocessing steps, in case users want to replicate or modify them. Note, in particular, the normalization module *normalizer.py* and the individual normalization grammars found in the *grammars/* directory.

7. Questions and feedback

Questions, comments and feedback about the NPSC are very welcome. We are also interested in corrections, modifications or derived resources (with a CC0 license or similar) that users make, which may be of interest to the speech technology community. To get in touch with us, use sprakbanken@nb.no.