National Library of Norway

NB Samtale

Norwegian conversation speech corpus

Version: 1.0

About the corpus

General introduction

NB Samtale is a speech corpus made by the Norwegian Language Bank at the National Library of Norway and contains orthographically transcribed speech from podcasts and recordings of live events at the National Library. The corpus is intended as an open source dataset for Automatic Speech Recognition (ASR) development, and is specifically aimed at improving ASR systems' handle on conversational speech.

The corpus consists of 12 080 segments, a total of 24 hours transcribed speech from 69 speakers. The corpus ensures both gender and dialect variation, and speakers from five broad dialect areas are represented. Both *bokmål* and *nynorsk* transcriptions are present in the corpus, with *nynorsk* making up approximately 25% of the transcriptions.

Licence

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

The audio material

The corpus consists of unplanned, conversational speech between two or more persons, and as such, include typical features of conversations, like overlapping speech and turn taking signals. The speech is unplanned in that it is not read from a script, nor a well-rehearsed presentation. Rather, the speakers speak freely in a conversation with others. That being said, one must assume that the speakers partake with a certain level of preparedness and many of the recordings follow an interview format.

The audio is collected from podcasts we have been permitted to share openly – namely *50 forskere* from UiT and *Trondheim kommunes podkast* from Trondheim municipality – as well as some of The National Library's own recordings of live events. The podcasts are studio recordings, while the National Library events take place in rooms and reception halls at the National Library, sometimes in front of an audience.

Gender and dialects

The recordings were for the most part selected based on the gender and dialect of the speakers to ensure gender balance and broad dialectal representation. The corpus has a

near 50/50 divide between male and female speakers (male 54%, female 46%). The Norwegian dialects have been divided into five broad dialect areas that are all represented in the corpus. However, Eastern Norwegian has the greatest representation at about 50% speaker time, while the other areas fall between 8% and 20% speaker time each.

The five dialect areas (and the counties they include) are:

- Eastern Norway (Østlandet): Agder, Innlandet, Oslo, Vestfold og Telemark, Viken
- Southwest Norway (Sørvestlandet): Rogaland
- Western Norway (Vestlandet): Møre og Romsdal, Vestland
- Central Norway (Midt-Norge): Trøndelag
- Northern Norway (Nord-Norge): Nordland, Troms og Finnmark

Speaker metadata and speaker ID

The recordings were segmented and transcribed in the transcription software ELAN. ELAN allows for multiple transcription tiers per file, and segments in the different tiers can overlap.

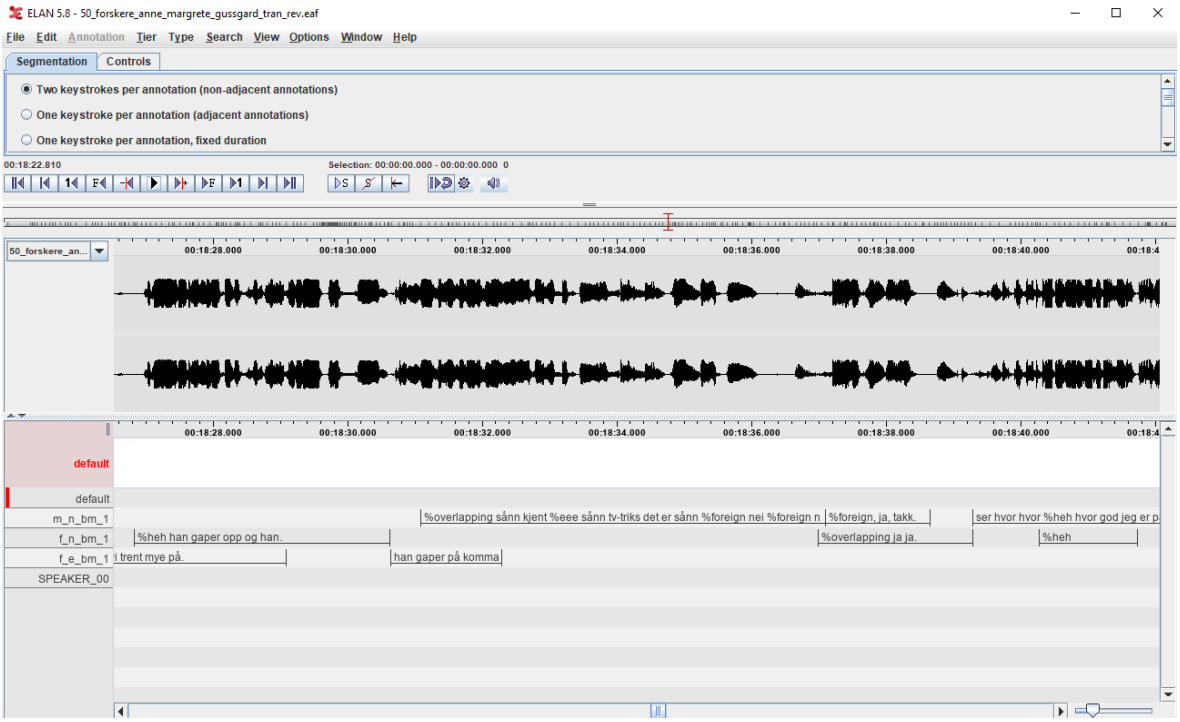

Figure 1: ELAN, segmentation window

All speakers in a file had separate transcription tiers and were named using speaker codes that denote metadata about the speaker. The metadata included the speakers' gender, dialect and the written standard of the transcription.The possible values are given in the overview below.

gender	dialect area	written norm
f (female) m (male)	e (east) n (north) sw (southwest) t (central) w (west)	bm (bokmål) nn (nynorsk)

Table 1: Overview of speaker code values

Some speakers feature in more than one file and there are 69 unique speakers in the corpus. The *metadata.jsonl* file has information about the unique speaker ID, gender, dialect and written standard for each segment in the corpus.

Many of the files also have a transcription tier for music and noise. This tier is given the ID P0 in the corpus and marks segments with music and non-human noise, such as applause.

Files

The file package contains a data directory with all the segments as individual WAV-files, as well as a jsonl-file with metadata and transcriptions. The WAV-files have a bit depth of 16 and a sample rate of 16kHz.

The file package also includes the file *filestructure.txt* and a *README.md* file. An overview of the file package structure is shown below:

```
.
├── data
│   ├── background_noise
│   ├── dev
│   │   ├── bm
│   │   │   ├── 50f-11_0006030-0014432.wav
│   │   │   ├── ...
│   │   │   ├── nb-1_0100800-0104784.wav
│   │   │   ├── ...
│   │   │   └── tr-31_0019850-0022397.wav
│   │   └── nn
│   ├── test
│   │   ├── bm
│   │   └── nn
│   └── train
│     ├── bm
│     └── nn
├── filestructure.txt
├── metadata.jsonl
└── README.md
```

Data directory

The data directory contains the segments split into a train, dev and test set for both *bokmål* and *nynorsk* transcriptions. The individual segments have names made up of the directory path (e.g. *data/train/nn/*) followed by the source file ID and segment ID (e.g. *nb-1_0100800-0104784.wav*).

The source file ID refers to the original file the segment is from with an abbreviation and number, i.e *nb-1*. The original files are numbered from 1 to 35, and additionally given a code denoting what podcast/series it is from: *nb* (National Library event), *tr* (*Trondheim kommunes podkast*), or *50f* (*50 forskere*).

The segment ID is its timestamp, that is, the start and end times of the segment in the source file. The combination of the source file ID and segment ID makes each segment name unique.

Most segments are shorter than 20 seconds and no segments are longer than 30 seconds.

Data split

The data has been split on three parameters: source type, gender and dialect. Gender and dialect naturally refers to the gender and dialect of the speakers. The data has not been split on speaker ID to avoid speaker overlap in the various sets because this proved impossible while still maintaining a decent distribution of the other parameters, especially dialect variation.

The source type refers to whether the source material is one of the two podcasts (*50f, tr*) or a National Library live event. The two types have different features. The podcasts are overall good quality studio recordings with little background noise, echo and such. The live events are recorded in rooms or reception halls at the National Library and have more background noise, echo and inconsistent audio quality. Many also have a live audience.

metadata.jsonl

The metadata file holds the information for each individual segment.

- source_file_id: original file the segment appears in
- segment_id: segment timestamp
- segment_order: order of segment in the original file.
- duration: duration of segment in seconds.
- overlap_previous: whether the beginning of the segment overlaps with the previous segment.
    - True or False
- overlap_next: whether the end of the segment overlaps with the next segment.
    - True or False
- speaker_id: speaker ID for the speaker transcribed in the segment.
- gender: speaker's gender
    - female (*f*) or male (*m*).
- dialect: speaker's dialect

    - *e* (east), *n* (north), *sw* (southwest), *t* (central) or *w* (west)
- orthography: the written norm of the transcription
    - *bokmål* or *nynorsk*
- source_type: type of recording of original file
    - *podcast* or *live-event*
- file_name: segment directory
- transcription: transcription

An example of segment metadata as found in *metadata.jsonl* is shown below:

```
{'source_file_id': 'nb-1',
 'segment_id': '0008970-0013860',
 'segment_order': 0,
 'duration': 4.89,
 'overlap_previous': False,
 'overlap_next': False,
 'speaker_id': 'P36',
 'gender': 'm',
 'dialect': 'e',
 'orthography': 'bm',
 'source_type': 'live-event',
 'file_name': 'data/train/bm/nb-1_0008970-0013860.wav',
 'transcription': 'hallo og velkommen hit til Nasjonalbiblioteket.'}
```

Principles for transcription

The recordings were segmented and transcribed in the transcription software ELAN. The recordings were transcribed automatically using a Norwegian ASR system created by the AI-lab at the National Library of Norway. The speech was segmented and transcribed with speaker diarization, separating the speakers into separate transcription tiers. These segments and transcriptions were then manually corrected by a transcriber according to a set of guidelines. All the manual transcriptions were reviewed by a second person in order to avoid substantial discrepancies between transcribers. Finally all the transcriptions were spell-checked, and checked for any unwanted numbers or special characters.

The full set of guidelines for segmentation and transcription are given in Norwegian in the file *NB_Samtale_transcription_guidelines.pdf*, but a summary of the main principles follow below.

Segmentation

The speech is segmented into grammatical sentences as much as possible, however, unplanned speech is characterized by interruptions, restarts, and other types of speech disfluencies that make it difficult to segment the utterances into neat sentences. Consequently, many segments consist of grammatically incorrect sentences.

The segments are generally not longer than 20 seconds, and never longer than 30 seconds. Sentences that are longer than 30 seconds are split into multiple segments. The sentences

are divided into clauses or phrases, splitting at for instance a coordination (e.g. *og (and), men (but), som (that/which)*) to satisfy the 30 second limit.

Overlapping speech is transcribed with overlapping segments in the different speakers' transcription tiers.

Transcription

The transcriptions are orthographic and adhere to either the *bokmål* or *nynorsk* norm, however the transcriptions aim to capture the speech as faithfully as possible by also including transcriptions of interruptions, repetitions and non-word sounds.

Written standard and backslash notations

When the written standard allows for multiple transcriptions of a word, the one closest to what was uttered is chosen. To illustrate, *bokmål* allows for both *-a* and *-en* endings on definite singular feminine nouns (i.e. *boka/boken*). If the speaker says /boka/ we transcribe *boka*. If the speaker says /boken/ we transcribe *boken*. Likewise, *nynorsk* allows for both *-a* and *-e* endings on infinitives (i.e. *kjøpa/kjøpe*). If the speaker says /kjøpa/, *kjøpa* is transcribed, while if the speaker says /kjøpe/, *kjøpe* is chosen.

Whenever a speaker uses a word, or has a dialectal pronunciation, that deviates from the respective written standard, we use a backslash notation with a close representation of what was uttered before the backslash and the dictionary form of the word after the backslash. For example, the word *beslutning* is not possible in *nynorsk*. The *nynorsk* equivalent is *avgjerd*. If a speaker transcribed in *nynorsk* utters the word /beslutning/, it is transcribed *beslutning\avgjerd*. Similarly, /mykje/ has a pronunciation that deviates from the *bokmål* standard *mye*. If the speaker is transcribed in *bokmål*, we use a backslash notation *mykje\mye*.

The threshold for using backslash notations is higher for function words. Function words feature more frequently in language, and past experience shows that the frequency with which they occur enable ASR systems to map dialectal pronunciations to standard spellings. Moreover, the transcription work is made more efficient by cutting down on backslash notations for this group of frequent words. Hence, function words only get a backslash notation when the pronunciation and standard spelling deviate significantly, e.g. *korsen\hvordan*.

There are 343 unique backslash notations in the corpus, listed in the file *backslash_notations.txt*.

Numbers, punctuation and capital letters

The transcriptions are in the spoken domain and do not contain digits, nor capital letters at beginning of sentence, and very little punctuation.

- All numbers are written out with letters. When a number is part of a word or name, the number is connected to the rest of the word with a hyphen.

- The transcriptions do not have capital letters at beginning of sentence, but capital letters are used in names, titles, acronyms and the like.
- Capital letters are also used when letters are spelled out.
- All sentences end with a period, or with a question mark or exclamation mark where appropriate.
- The transcriptions have commas in the conventional places, mostly at coordinators, embedded clauses, and in lists. The use of commas is in many cases optional, and we have included fewer commas rather than more.
- If a long sentence has been split into multiple segments to uphold the 30 second limit, and a period is not appropriate at the end of the segment, a comma is used instead.
- The transcriptions do not contain other punctuation like parentheses, colons, slashes, quotation marks, etc.

Acronyms

Acronyms that are pronounced as words are written out in the conventional way (e.g. /nato/ → *NATO*). Acronyms that are pronounced by spelling out each letter (e.g. /en ær kå/ → *NRK*) are trickier to transcribe. The pronunciation deviates from the spelling in that the letters are spelled out and, as per the guidelines, should be written out with individual capital letters (/en ær kå/ → *N R K*). Yet, it is beneficial to map the pronunciation to the conventional spelling (/en ær kå/ → *NRK*). Due to the discrepancy between pronunciation and spelling, these acronyms are also transcribed with backslash notations that capture both a close representation of what was uttered as well as the conventional spelling: /en ær kå/ → *N_R_K\NRK* (underscores are used in place of spaces before the backslash). This also ensures that the acronyms are marked and can easily be found in the corpus.

Interruptions and non-words

The transcriptions are an accurate reflection of the speech and include interrupted words, hesitations and other sounds.

- Interruptions are marked with a £ at the end.
- Non-word sounds have been described using three-letter-labels that begin with %
    - oral hesitation: *%eee*
    - nasal hesitation: *%mmm*
    - oral + nasal hesitation: *%emm*
    - laughter: *%heh*
    - coughs and the like: *%qqq*
    - tuts, smaks and the like: *%ttt*
    - gasps and the like: *%hhh*
- Breathing in general is not transcribed
- Words, phrases and sentences in a foreign language are not transcribed. The label *%foreign* is used instead. Names and titles in foreign languages, however, are transcribed.
- Unintelligible speech is marked with the label *%unint*.
- When overlapping speech causes the speech to be unintelligible, it is marked with the label *%overlapping*.

- If the transcriber hears what is said, but is unfamiliar with the word's meaning and spelling and unsuccessful in finding the correct transcription, the label *%unk* for *unknown* is transcribed.
- Music and non-human sounds, such as applause or other loud background noises, are transcribed in the *music* tier. Music is marked with the label *%music*, and other noises with the label *%noise*.

Questions and feedback

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