Nasjonalbiblioteket

Norwegian Parliamentary Speech Corpus

The Norwegian Language Bank

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 2019-2020. 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

A beta version of the NPSC is launched in October 2020. The following audio files, with transcriptions and metadata, are included in this release

Filename	Duration in hours
20170207-095506.wav	1:14
20170208-095517.wav	2:29
20170209-095509.wav	1:40
20170215-095508.wav	3:55
20170216-095707.wav	1:48
20170314-095510.wav	3:50
20170403-115516.wav	3:37
20170405-095525.wav	1:58
20170419-095601.wav	1:38
20170503-095709.wav	3:12
20170510-095755.wav	4:00
20170613-085456.wav	6:10
20170615-085510.wav	6:10
20171007-125517.wav	1:19
20171018-095533.wav	2:45
20171211-095523.wav	6:10
20171219-095511.wav	6:10
20180201-095509.wav	1:40
Total:	58:49

Table 1: File names

Note that the total duration of the sentence-segmented files is somewhat shorter than the total in the table below, as pauses are not included.

A stable version of the NPSC will be launched in 2021. This version will be substantially larger than the beta version. For the stable version to be as good as possible, feedback on the beta version and suggestions for improvement are very valuable. You can reach us at sprakbanken@nb.no.

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

Transcription pipeline

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

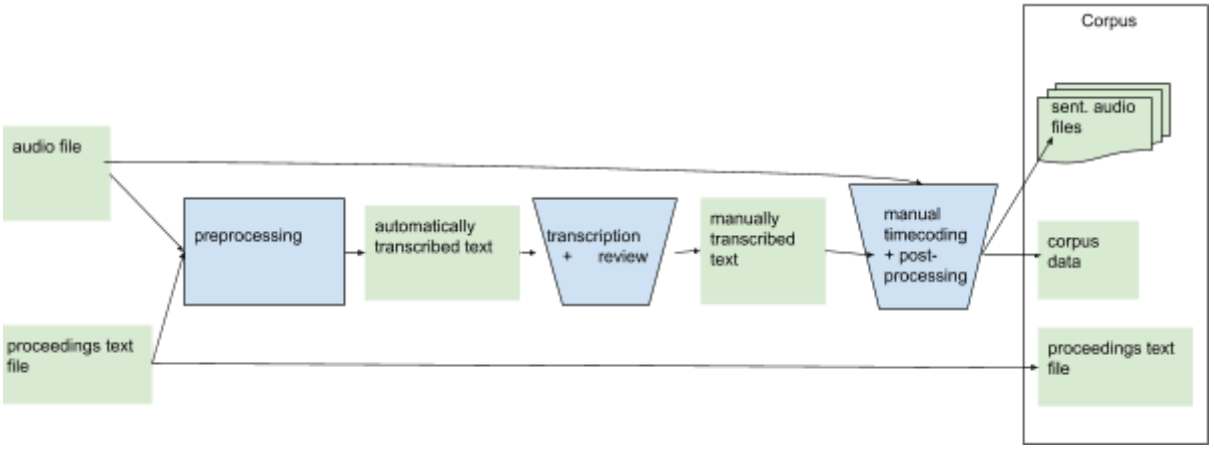

Figure 1: Transcription pipeline

[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.

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.

Description of files and directories

Naming conventions and directory structure

All files in the corpus are uniquely named, so there 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/*. This directory contains three files, each named with the date and the start time of the recording and a file signature: *yyyymmdd-hhmmss.**, e.g. *20170207-095506.**, 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 exactly six minutes and ten seconds, which will usually be in the middle of a sentence. The last sentence of the audio file and the transcription will therefore be incomplete in these cases.
- *20170207-095506_corpusdata.json* contains the transcription and information about the speakers, as described in the next subsection.
- *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*.

The content of the corpusdata file

The corpusdata json file contains the transcriptions and other relevant data. On the top level, the corpusdata file is a dictionary with the following keys: *meeting_id, meeting_date,*

audio_file, proceedings_file, duration_ms, transcriber_id, reviewer_id, speaker_data, sentence_data, token_data. All the keys except for the last three have string or integer values giving relevant metadata about the transcribed meeting and the files: *meeting_id* is the identifier of the meeting in our internal database, *meeting_date* gives the date of the meeting, e.g. *20170207*, *audio_file* and *proceedings_file* render the names of the relevant .ref and .wav files, *duration_ms* contains the duration of the audio file in milliseconds, *transcriber_id* and *reviewer_id* contain the identifiers of the transcriber and reviewer respectively in our internal database. *speaker_data, sentence_data* and *token_data* contain lists of dictionaries with information about speaker, sentences and tokens (words) respectively. The keys of these dictionaries are described in the following:

- *speaker_data* contains a list of dictionaries for each speaker in the transcription. These contain the following key-value pairs:
    - *speaker_id*: an identifier for the speaker, which corresponds to the *speaker_id* in the *sentence_data* dictionaries.
    - *speaker_name*: the full name of the speaker, or "*unknown"*, if the speaker is unknown
    - *speaker_URI*: the URI of the speaker in Wikidata, if it exists, else *null*.
    - *date_of_birth:* the date of birth of the speaker according to Wikidata, if it is found there, else *null*. The format is *yyyy-mm-dd*.
    - *place_of_birth:* the name of the place of birth according to Wikidata, if it is found there, else *null*.
    - *pob_URI:* the URI of the place of birth in Wikidata, if it is found there, else *null*.
    - *pob_county*: the county of the place of birth according to Wikidata, if it is found there, else *null*.
    - *pob_region*: the region of the place of birth according to Wikidata, if it is found there, else *null*.
    - *electoral_district*: the electoral district of the speaker according to Wikidata, if it is found there, else *null*. Note that the electoral district is only relevant to MPs and deputy MPs, not to members of government and other speakers.
    - *ed_URI:* the URI of the electoral district in Wikidata, if it is found there, else *null*.
    - *gender:* the gender of the speaker, unless the speaker is unknown, in which case the value is *null*.
    - *written_standard:* The language code of the written standard the speaker has chosen, either *nb-NO* or *nn-NO*.
- *sentence_data* contains a list of dictionaries for each sentence of the transcription. These dictionaries have the following key-value pairs:
    - *sentence_id* is the identifier of the sentence. This corresponds to *sentence_id* in the *token_data* dictionaries.
    - *sentence_text* is a string representation of the sentence, including interruptions, repetitions and hesitations, written as <*ee*> for vocal hesitations, <*mm*> for nasal hesitations and <*qq*> for other, non-verbal vocalized sounds. It can also contain the string <*INAUDIBLE*> indicating inaudible or overlapping speech. The string representation does not include any

annotations for non-standard forms etc. *sentence_text* is simply a concatenation, separated by whitespaces, of the *token_text* values of the tokens of the dictionaries in *token_data*.[2] As will be explained below, *token_text* may contain non-standard words and interrupted words, which will also be rendered in *sentence_text*. Users who want to exclude or mark in some way non-standard words or interrupted words, cannot rely solely on *sentence_text*, but need to fetch that information from the tokens.

- *sentence_audio_file* contains the filename of the sentence wav file
- *sentence_start_time_ms* contains the start time in milliseconds of the sentence in the full audio file.
- *sentence_end_time* contains the end time in milliseconds of the sentence in the full audio file.
- *speaker_id* contains the identifier of the speaker and corresponds to the *speaker_id* field in the *speaker_data* dictionary.
- *sentence_language_code* contains the language code of the sentence, for the most part *nb-NO and nn-NO,* but also occasionally *en-GB* and possibly other languages. The sentence language is the same as the code in the *chosen_language* for the speaker of the sentence in *speaker_data*, unless all of the tokens of the sentence have been annotated explicitly with a different language code. If all of the tokens have been annotated with a different language code, this language code will be assigned to *sentence_language_code*. If only some of the tokens are annotated with a different language code, the chosen language of the speaker takes precedence.

- *token_data* contains a list of dictionaries for each token of the transcription. Since the transcription does not include punctuation, token boundaries are always whitespace or sentence boundaries. However, there are certain non-standard multiword expressions (MWEs), e.g. “i henhold til”, ‘according to’ (non-standard in Nynorsk, but not in Bokmål), which have been treated as one token.
  - *token_id* contains an identifier for the token
  - *token_text* contains a string representation of the token. In cases where nonstandard MWEs are one token, words are separated by an underscore. Hesitations are rendered as *<ee>, <mm>* or *<qq>* (see *sentence_text* above). Inaudible or overlapping speech is rendered as *<INAUDIBLE>*. Incomplete or interrupted words have a string representing the partial word or the sounds produced.
  - *nonstandard_spelling* contains Boolean values: *true* if *token_text* is spelled in a way which is not compliant with the official standard of the language of given in *token_language_code*, *talse* if it is compliant. Criteria for choosing a non-standard spelling are given in the next section.

[2] Note, however, that there are a few cases where words in multiword expressions are tokenized together. In these cases, there will be an underscore between the words in *token_text*, cf. <ref>. In *sentence_text,* there will be a whitespace between the words of the multiword expression.

- *standardized_form.* If *nonstandard_spelling* is true, a standardized equivalent with the same meaning is given in this field. If the standardized form is an MWE, words are separated by an underscore.
- *special_status* is used for tokens that does 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*.
- *token_language_code* is set as equal to *chosen_language* for the speaker in *speaker_data*, unless the transcriber has explicitly annotated the token with a different language code, which could be *nn-NO, nb-NO, en-GB*, and possibly others.
- *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 *true* in this field. All other words have the value *false*.
- *sentence_id* is the identifier of the sentence to which the token belongs, and corresponds to *sentence_id* in *sentence_data*

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. These guidelines are written by two 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 its 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).

Transcription domain

Orthographic transcriptions are sometimes divided into *spoken* and *written domain*. In *written domain* transcriptions, words and phrases are rendered as they would typically be in a written text: numbers are written with digits, e.g. *62*; dates and times are written in a standardized format, e.g *9.10. 2020 kl. 16.50*; common abbreviations are used, e.g. *f.eks.,* 'e.g.'. This typically corresponds to what an ASR system produces. In *spoken domain* transcription, the pronunciation of words and phrases are rendered as accurately as the written standard allows: numbers are written as words, e.g. *sekstito*, 'sixty-two'; dates are written as they are pronounced, e.g. *niende oktober tjuetjue*, 'October ninth, twenty twenty'; abbreviations are not used unless the speaker explicitly says the abbreviated form of the expression.

In the NPSC we have chosen a spoken domain transcription in order to be as faithful to the spoken word as possible (principle 3). We are looking into ways of converting the transcriptions to written domain at a later stage.

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.[3]

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 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

[3] 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.

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*. Words in the following classes are considered function words: pronouns, determiners, prepositions, complementizers, conjunctions, auxiliary verbs, interrogative adverbs, negation and some temporal and locative expressions meaning then and now and here and there. Note that the marking of function words isn't entirely consistent, as it is difficult to make precise criteria for when a function word deviates significantly from the standard.

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*. 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 *sentence_text*, 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 explicitly 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 is transcribed as <*qq*>. These are also marked with *special_status: HESITATION* in *token_list*.

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*.

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, which are added to *speaker_data*. If this information is not found in Wikidata, the relevant fields are null in *speaker_data*. Based on the regional data, it is possible 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. When a speaker is not in Wikidata, we have made an effort to add them there, including the relevant metadata when we have been able to find it.