MLLP | Machine Learning and Language Processing

VRAIN Valencian Research Institute for Artificial Intelligence

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

Europarl-ASR

Manual revision guidelines

https://www.mllp.upv.es/europarl-asr

2 April 2021

1 Goals

The main goals that we want to achieve with this revision process are the following:

1. To get correct and precise start and end times of each speech/intervention.
2. To generate raw/verbatim/exact transcriptions of each speech/intervention, that could be suitable for ASR system development and evaluation.

2 Preliminary considerations

- Audio files corresponding to an isolated speech/intervention are trimmed from the original video files of the European Parliament (EP) session, using the official start and end times published at the EP website. **These time marks are often inaccurate, and, as a result**:

  - **A fragment of previous/latter speeches might be included at the beginning/end of the trimmed audio file**: please do NOT transcribe those fragments. Note that the correct, exact start and end times of the whole speech will be inferred from the first and last revised transcription segments, and these will be used afterwards to re-trim audio clips with much more accuracy (achieve goal no. 1).
  - **The whole speech is not contained in the audio file**: the actual speech could have started before the audio file starts, or could have ended after the audio file ends. In this case, please, state this issue in the comments section of the provided CSV file, let the revision coordinator know, and skip the file.

  Therefore, **the very first step of the revision process** is to check whether the whole speech is fully included in the trimmed audio file or not. If not, please do not review the file, state this issue in the "Comments" column of the CSV file, and go to the next file. Do this also if the **interpreter's (dubbed) speech overlaps with the main speaker's speech** (as the whole speech is not fully included in the audio as well).

- Speech transcripts available at the EP website **are not verbatim, but *free* transcriptions**, that:

  - may omit words or phrases, i.e., *non-transcribed speech.*
  - may add unuttered words or phrases, i.e., *injected text.*
  - may correct words or phrases to make them linguistically correct, i.e., *amended transcription.*

  - may use synonyms, or even rephrase and/or summarize the actual speech, i.e., *rephrased transcription.*

  These transcripts should be post-edited in order to convert them into verbatim ones, considering the transcription guidelines described in Section 3. This is to achieve our goal no. 2.

- A reduced set of speech transcripts have been previously filtered out by applying a speech data filtering criterion based on the Character Error Rate (CER), computed over automatic transcriptions of the speeches, generated by an ASR system, and using official transcripts as reference. Speeches showing CER values higher than a certain threshold (manually set after empirical observation of the data) were removed. This is to ensure that bad audio-transcript mappings are not present in the final dataset. However, if you consider that a particular speech should be filtered out as well, please inform the revision coordinator.

- The revision will consist of post-editing subtitles (segments) generated automatically after running a forced-alignment process between the trimmed audio file and its corresponding official (non-verbatim) speech transcript. Therefore, there might be alignment errors, specially:

  - at the beginning and at the end of the audio file and transcript
  - with *non-transcribed speech*,
  - with *injected text.*

  These alignment errors should be corrected by adjusting segment start and end times, when necessary.

- With the exception of the first and last segments of the whole transcription (remember our goal no. 1), the inner text segmentation is not really important (i.e., it is not necessary to define segments that fit syntactical units or other recognizable linguistic units). Any inner text segmentation will work, provided that start and end times of every segment are well adjusted to the audio.

- Segments should not include long passages of non-speech sounds, unintelligible speech, etc., as we plan to generate reference `.stm` files.

- In case of doubt about how to proceed with specific cases, please contact the revision coordinator.

3 Transcription guidelines

Always transcribe what was said:

- We are looking for raw, verbatim transcriptions (not subtitles).
- Transcribe as phonetically as possible, using valid words (except incorrect pronunciation that changes meaning, see “Speech disfluencies”).
- Always transcribe what is spoken, do not change or correct what the speaker says (except incorrect pronunciation that changes meaning).
    - Example: (saying) “*it’s me!*” → “*it’s me!*”. Do NOT transcribe as “*it is me!*”.
- Do not correct grammar.
- Include punctuation marks and capital letters, when needed.
- If a punctuation mark is spoken, it should be transcribed as a word.
- When the speaker corrects themselves: transcribe what was said.
- Do not use abbreviations, e.g., “*Mr.*” → “*Mister*”
- Use standard (ASCII) punctuation mark symbols.

Letters and spelled out words:

- Each letter will be individually typed and uppercased, separated from the other letters by a space.
    - Example: “*my name is Smith S M I T H*”
- For acronyms (e.g., “*USA*”), the individual letters will not be separated by spaces. However, if the speaker pronounces all the words (e.g., “*United States of America*”), keep the expanded text as transcription.

Numbers, digits, ordinals, variable names, equations:

- Numbers, digits and ordinals will be transcribed as words.
    - Example: “*four hundred eighteen, forty-fourth, two thousand and two*”
- Decimals are transcribed as words.
    - Example: “*two point six*”

- Variables are transcribed as words.
  - Example: "*epsilon, lambda, ...*"
- Equations must be written out in words as they are spoken.
  - Example: "*P equals N over V times R T*"
- Zero ("0"), when pronounced "oh", must be transcribed as "oh".
- Further examples: "*three squared plus two, seven point three minus three point eight*".

Non-speech events:

- Do NOT transcribe them. Examples:
  - Speaker noise (unvoiced sounds): lip, smack, cough, etc.
  - Non-stationary noise: door slam, window, etc.
  - Music.
  - Chatter and noise.

Speech disfluencies:

- **Hesitations**: label them according to language-specific instructions (see Section 3.1).
- **Incorrect pronunciation**: a speaker incorrectly utters a word from a phonetic perspective, causing a different word/sentence meaning. Transcribe with the correct word (the one they wanted to say).
  - Example: (saying) "*beer* in mind" → (transcript) "*bear* in mind" (from the context, they actually wanted to say "*bear*").
  - But not: (saying) "has to be *uphold*" → (transcript) "has to be *upheld*" (as the meaning doesn't change, it's just a grammatical error, keep "*uphold*")
- **Foreign words**: for words, titles, etc. in foreign languages that are pronounced differently from the main spoken language pronunciation rules, they are to be transcribed in the foreign language.
  - Example: "*there was a coup d'êtat in ...*", "*Thank you, Monsieur Président*".
- **Repetition**: the speaker, unintentionally, repeats a word or expression. It is transcribed as it is pronounced.

- **Word cut-offs**: imagine the speaker utters only half a word. This half-word is to be transcribed as spoken, with a hyphen at the end.
  - Example: "*From the bot- uh the bottom of my heart*"

Other situations:

- **Incomprehensible passages**:
  - Short ones (i.e., 1-3 words, isolated or within a comprehensible passage): use `<UNK>` for the unknown words.
  - Moderate-long to long ones: do not transcribe them (leave them out of any segment).
- **Overlapping speech**: transcribe only what the main speaker says.
- **Foreign language sentences or passages**: do not transcribe them (leave them out of any segment).

3.1 Language-specific instructions

3.1.1 English

- Label hesitation sounds as "uh" or "um".