Underrepresented speech dataset from open data: case study on the Romanian language
(USPDATRO)

Annotation Guidelines

Table of Contents

Content identification and download 3

File based metadata 5

Creating aligned text 5

Text format 8

Content identification and download

Each of the targeted multimedia platforms offers search functionality that allows retrieval of multimedia content specifically marked as being available under an open license. In addition, the USPDATRO project is focused on underrepresented voice types (particularly young adults, old adults and women, while other cases may be identified as well). Surveyed multimedia platforms (see USPDATRO Final Report) do not allow for explicitly searching for such voice characteristics. Therefore, for proper identification of content, specific search keywords or key-phrases must be used. Furthermore, these must take into account the focus on the Romanian language, therefore keywords employed must be in the Romanian language and specific to this language (without similar words being present in other languages).

Proposed search words and phrases:

Keywords	**Target group**
liceu elevii te învață sugestii pentru scoala povesti pentru copii	Young people
pentru tineri sfaturi pentru tineri	Older people
emisiune pentru femei	Women
emisiune radio	Generic

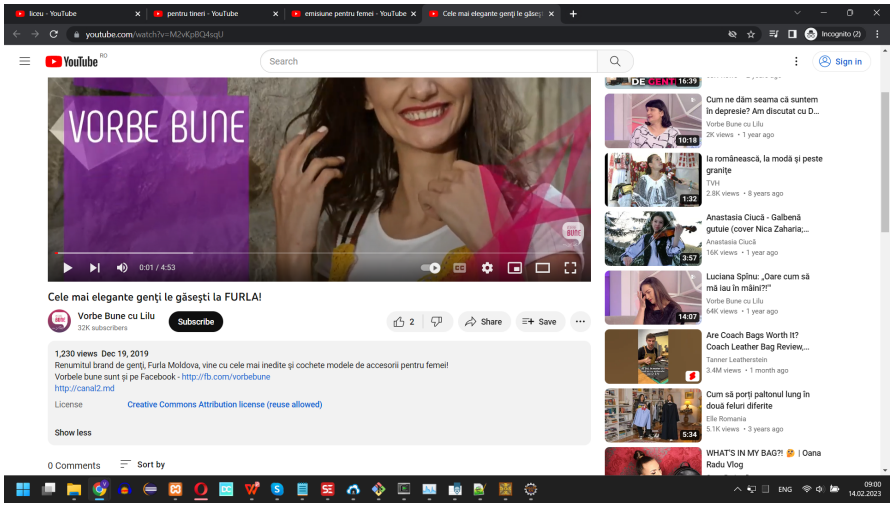
Once suitable candidate multimedia recordings have been identified, it is important to double check in order to make sure that the uploader has the rights to give the content under the specified license. This usually involves confirming that the uploader is the content producer or has the rights to distribute it. There are two possibilities:

a) The uploader is an individual: in this case it is important to check that he/she is the actual content producer. This usually involves having many such multimedia recordings on their account (possibly some with other licenses).
b) The uploader is an organization: in this case the organization may be the content producer (this should clearly be indicated in the videos and is usually the case with televisions offering open content), or the organization should have proper rights to distribute the content (and this should somehow be explained in the information associated with the account).

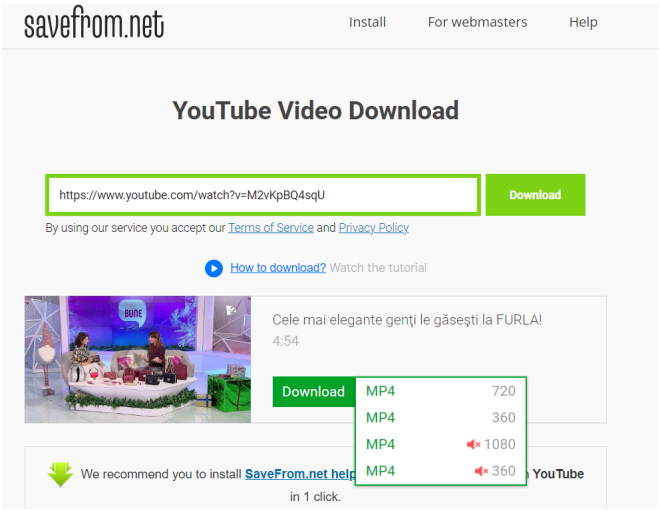
After clarifying the license, the content must be downloaded into a file. Since multimedia platforms usually host video content, the resulting file will be a video file. For the final corpus release this will be converted into an audio file. However for the following operations (file based metadata, subtitle generation and speaker metadata) it may be useful to keep the file as a video

file (video information may provide additional hints for constructing the metadata or clarifying the subtitles).

When downloading the file, it is important to also take a screenshot of the platform clearly showing the account and the license associated with the content. This will not be part of the final corpus release, but will allow confirming the associated license in case the content will be removed at a later time. Example:



Depending on the platform, the file can be downloaded automatically (from within the platform) or via 3rd party applications or websites. An example of such a website is https://en.savefrom.net/383/ , which allow downloading content from multiple multimedia platforms (for example YouTube download can be accessed here: https://en.savefrom.net/1-youtube-video-downloader-437/).Furthermore, when downloading the website allows specification of the file quality being generated. This allows reducing the space needed for storing the corpus during processing:



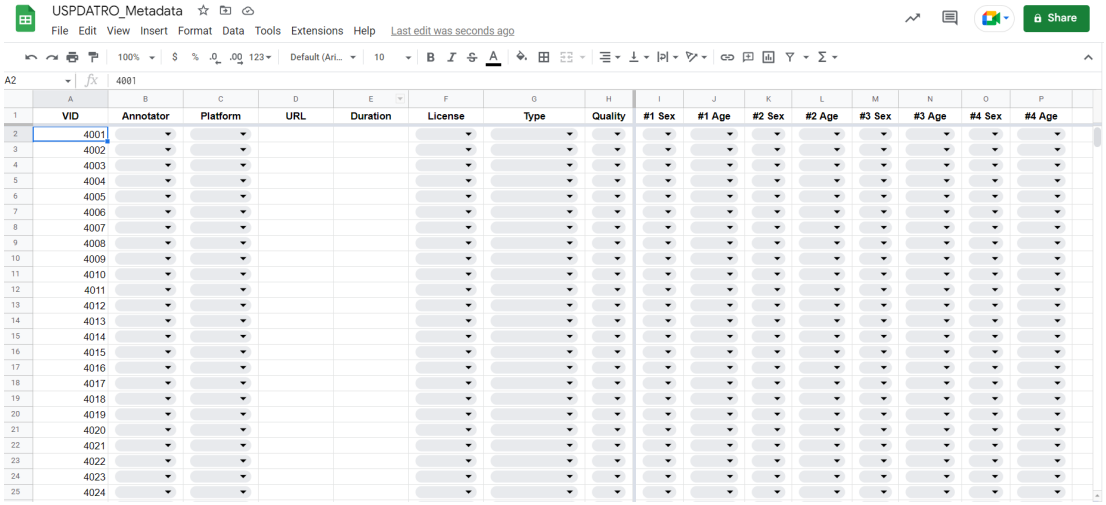
Since the end objective of the project is to produce a speech corpus, it is recommended to download the videos at the lowest possible resolution that still allows for generating the subtitles. For example, selecting the 360p video resolution, a 5 minute video is downloaded into a file of 21Mb in size.

File based metadata

The following metadata fields will be completed for each downloaded file:

- *Platform* : this will indicate the platform from which the content was downloaded
- *URL* : the URL from which the multimedia content is available
- *Duration* : this will be the time reported in the platform, in the format hh:mm:ss. This is the total duration of the file, which is usually less than the usable duration (the part containing relevant voices).
- *License* : one of the open licenses, as indicated in the platform by the content uploader
- *Type*: this indicates the type of speech: read or spontaneous
- *Quality*: MOS (Medium Opinion Score) quality index (5=Excellent, 4=Good, 3=Fair, 2=Poor, 1=Bad)
- *Speakers*: for each speaker the following information is provided:
    - *Gender*
    - *Age group*

For storing the metadata a Google Sheets document was setup with data validation rules considering dropdown and allowed data formats.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	VID	Annotator	Platform	URL	Duration	License	Type	Quality	#1 Sex	#1 Age	#2 Sex	#2 Age	#3 Sex	#3 Age	#4 Sex	#4 Age
2	4001															
3	4002															
4	4003															
5	4004															
6	4005															
7	4006															
8	4007															
9	4008															
10	4009															
11	4010															
12	4011															
13	4012															
14	4013															
15	4014															
16	4015															
17	4016															
18	4017															
19	4018															
20	4019															
21	4020															
22	4021															
23	4022															
24	4023															
25	4024															

Only Romanian speakers will be considered. If the file has music, non-Romanian speakers or other sounds, these will not be considered.

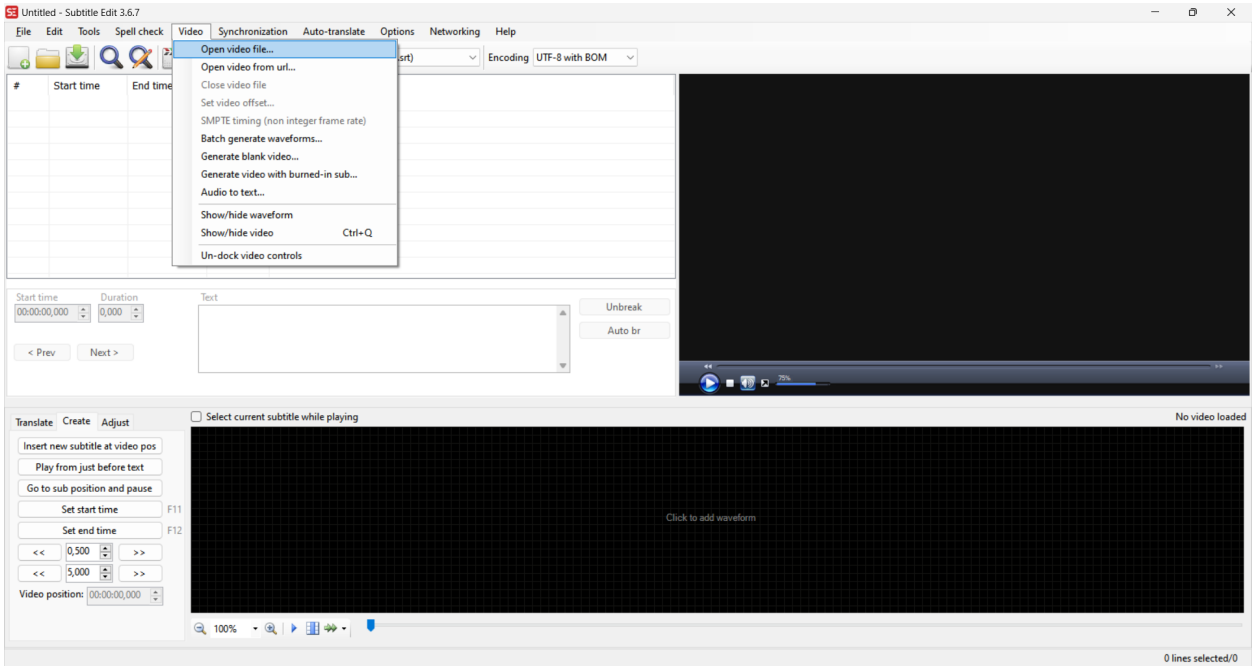
Creating aligned text

Aligned text with multimedia files is commonly known as subtitles. However, for the purposes of training deep learning algorithms able to process speech, the resulting text must be well aligned

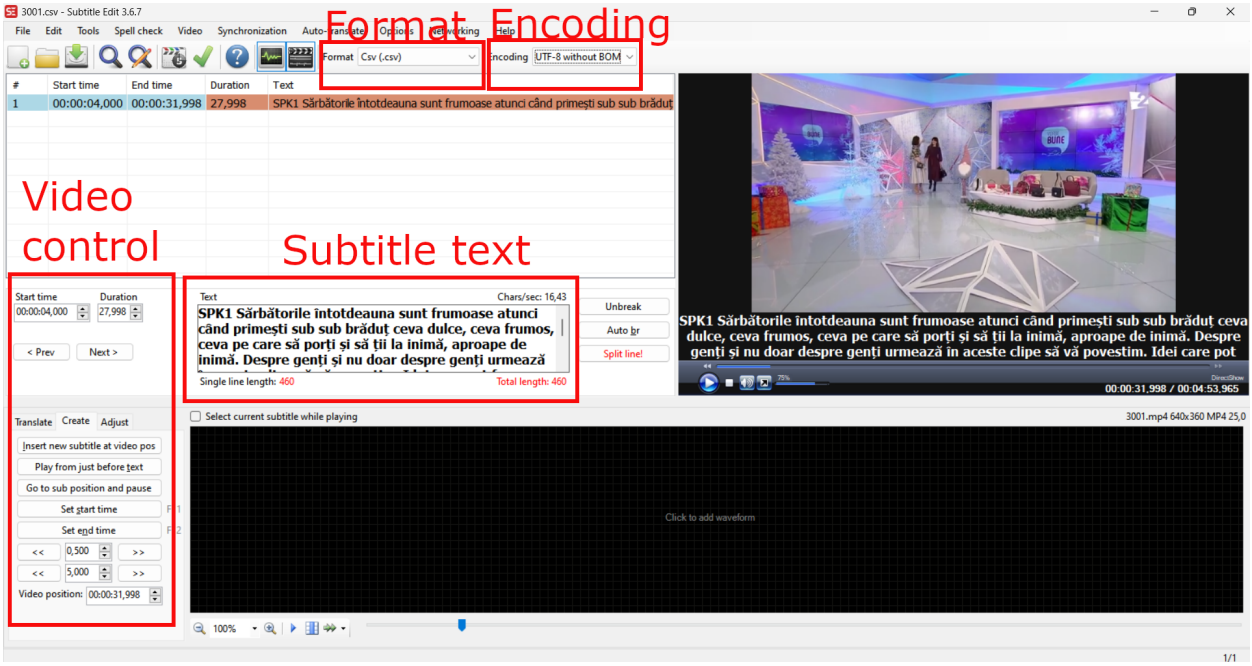
with the audio data. Furthermore, we want to explicitly indicate to which of the speakers a certain text belongs to.

For the project's purposes we will use Subtitle Edit, which is a free software under the GNU Public license: https://www.nikse.dk/ , https://github.com/SubtitleEdit/subtitleedit .

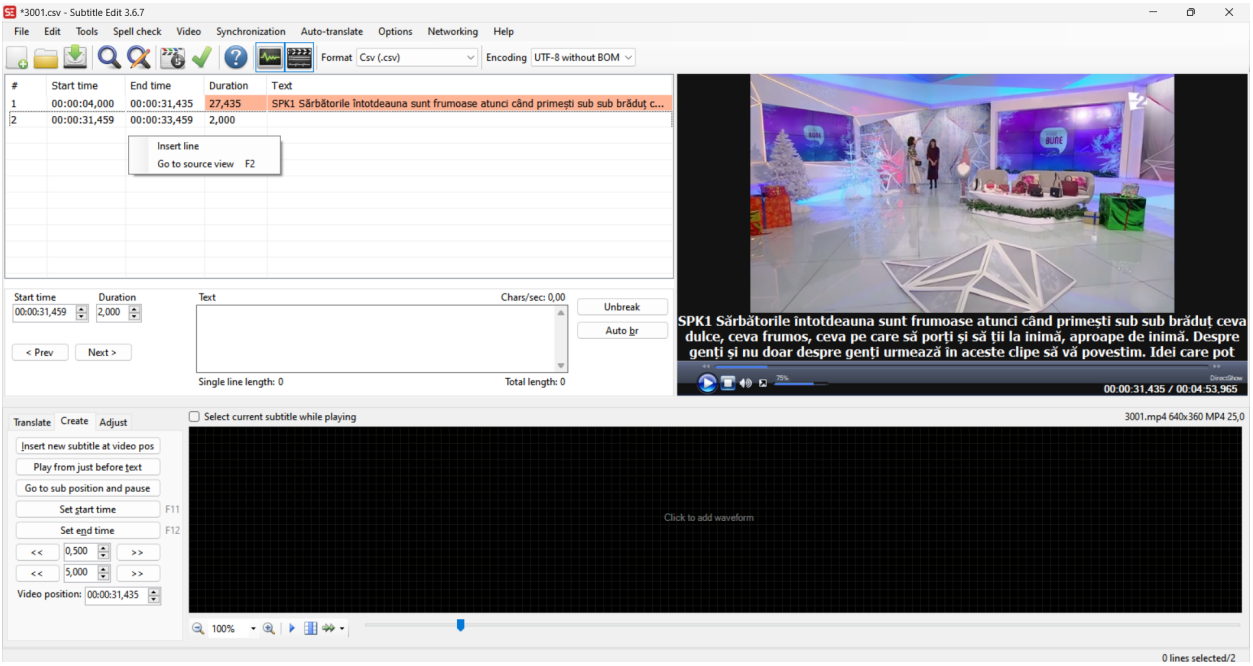
Given a new SubtitleEdit project, first the video file is opened:



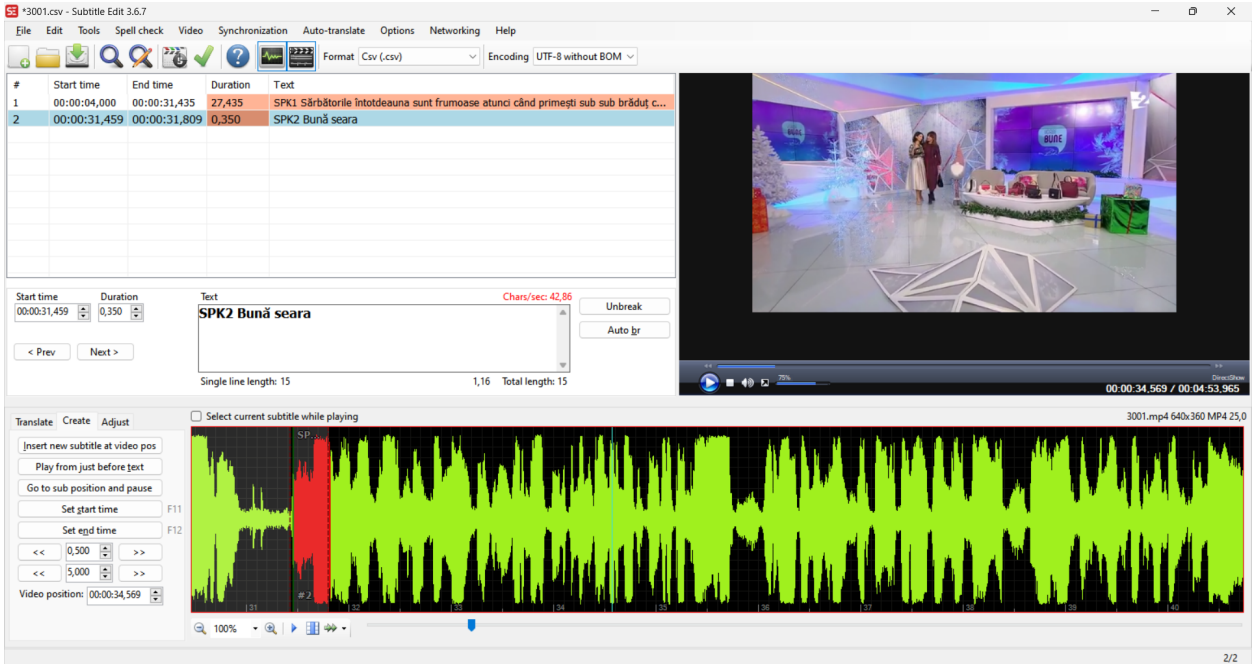
The text associated with the audio is entered in the center part of the screen. The controls from the bottom left of the screen allow setting the start/end position of the text and also allow skipping the video in small amounts of time in order to better identify the positions.



Adding a new text fragment can be accomplished by right clicking in the subtitles area and selecting “Insert line”:



With a click on lower part of the screen the waveform can be visualized in order to allow more precise alignments:



Because the editor is not aware of speaker changes, each text fragment must be prefixed with “SPK”+NUMBER. Example: “SPK2 Super”.

Text format

When selecting the appropriate “CSV” format and “UTF-8 without BOM” encoding, the resulting CSV file will contain the following information:

- Number: the text number within the file
- Start time in milliseconds
- End time in milliseconds
- Text: this is the actual text fragment surrounded in quotation marks.

```
Number;Start time in milliseconds;End time in milliseconds;"Text"
1;4000;31435,0793;"SPK1 Sărbătorile întotdeauna sunt frumoase atunci când primești sub sub brăduț ceva dulce, ceva frumos, ceva
2;31459,0793;31809,0793;"SPK2 Bună seara"
3;31845;39277,8549;"SPK1 din privința unei femei pentru un bărbat, pentru că de obicei bărbații cumpără cadouri și nu întotdeau
4;43390,0907;44776,6167;"SPK2 Mersi mult. Mersi mult."
5;44800,6167;53043,8625;"SPK1 Se apropie sărbătorile și noi avem idei de cadouri aici pe masă și în mâna mea. Asta-i de seară ș
6;105035,6856;113282,1021;"SPK1 Da. E e important. Știu că ați fost recent și la Milano fashion week... Ce te-a impresionat acc
7;113306,1021;199477,9977;"SPK1 Da. E minunat. Știu că ați avea și o nouă față."
8;234297,8859;237438,5329;"SPK1 Frumos. Știu că ne-ai adus astăzi și cadouri."
9;238675,6863;243112,3598;"SPK1 Eu tot timpul mă uit așa. Da cu ce vine un invitat de-al nostru. Oare are ceva pentru mine sau
10;268954,328;271744,9864;"SPK1 Așa-i. Și pentru telespactatorii noștri"
11;273441,4149;282277,5813;"SPK1 Așa că î stați aproape de noi. În curând vom posta dar pe mine în curând o să mă mai vedeți î
12;282277,5813;282654,5654;"SPK2 Super"
13;282767,4936;289903,1882;"SPK1 Și și și e tare cul. Și pentru fetiță. Pun lucrurile în ea. Mulțumim tare mult că ai reușit să
```

In order to allow association between the text and the corresponding multimedia file, the file name will be kept the same.