EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

USPDATRO – Underrepresented Speech Dataset from Open Data: Case Study on the Romanian Language

Authors	Vasile Păiș, Verginica Barbu Mititelu, Elena Irimia, Radu Ion, Dan Tufiș
Organisation	Research Institute for Artificial Intelligence, Romanian Academy

Dissemination level	Public
Date	DD-MM-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	USPDATRO – Underrepresented Speech Dataset from Open Data: Case Study on the Romanian Language
Authors	Vasile Păiș, Verginica Barbu Mititelu, Elena Irimia, Radu Ion, Dan Tufiș
Organisation	Research Institute for Artificial Intelligence, Romanian Academy
Type	Report
Number of pages	27
Status and version	Draft
Dissemination level	Public
Date of delivery	DD-MM-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AI4EU	AI4EU (EU project, 2019-2021)
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CRACKER	Cracking the Language Barrier (EU project, 2015–2017)
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DLE	Digital Language Equality
EC	European Commission

ECSPM	European Civil Society Platform for Multilingualism
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality
ELE2	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELEXIS	European Lexicographic Infrastructure
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
ELT	European Language Technology
EP	European Parliament
ERIC	European Research Infrastructure Consortium
ESCO	European Skills, Competences, Qualifications and Occupations classification
GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NCP	National Contact Point
NLP	Natural Language Processing
STOA	Science and Technology Options Assessment

Abstract

The USPDATRO project's goal was to study the usability of open data for building speech datasets for types of voices that are usually missing from or are underrepresented in existing speech datasets. A case study on the Romanian language was conducted, also investigating the possibility of applying the same methodology to other languages. Existing multimedia platforms were investigated in order to discover open multimedia data (available under open licenses). The report covers the platforms, types of media, percent of usable voices in a collected Romanian data sample, types of open licenses, types of underrepresented voices (including children, young people, older people, women, etc.), percent of underrepresented voices. A pilot dataset of Romanian underrepresented voices aligned with the corresponding textual representation was constructed and released.

1. Introduction

The project consisted of 4 activities, following the project proposal. The activities implementation and results will be detailed in the following sections of this report. The overall goal of the project was the identification of speech publicly available under open licenses (public domain or Creative Commons attribution), usable for building improved ASR systems in lower resourced languages. We focused primarily on the Romanian language, while maintaining the methodology applicable for any other language. Furthermore, throughout the project we paid special consideration to voices that are usually underrepresented in traditional speech datasets, such as children, old people, women, and non-natives.

2. Activity 1. Identification of multimedia platforms with open content

This activity involved checking the established multimedia platforms for available licensing options. We were primarily interested in the availability of open licenses. Furthermore, the search features of these platforms were investigated in order to assess their usability for the project's objectives. This includes searches based on language, open license, and features allowing the identification of underrepresented speech types. The activity covered the following platforms, as detailed below: YouTube, Vimeo, TikTok, SoundCloud and LinguaLibre.

2.1. YouTube

YouTube[1] is a very popular multimedia sharing platform. According to the "About"[2] page, their "mission is to give everyone a voice and show them the world". The platform allows users to upload either long videos or short ones (referred to as "YouTube shorts"). In this context, users can be either individuals or organizations who own a YouTube account.

The "Copyright"[3] page gives indications that "Creators should only upload videos that they have made or that they are authorized to use. That means they should not upload videos they did not make, or use content in their videos that someone else owns the copyright to, such as music tracks, snippets of copyrighted programs, or videos made by other users, without necessary authorizations". This makes the users responsible for the content they provide.

[1] https://www.youtube.com/
[2] https://about.youtube/
[3] https://www.youtube.com/howyoutubeworks/policies/copyright/

In the YouTube platform, Creative Commons licenses give a standard way for content creators to grant someone else permission to use their work[4]. The default license associated with any new content is the YouTube license. Any user wanting to make their own content available under Creative Commons license must select this license explicitly. Prior to September 2021, the option to associate a Creative Commons license was available in the view attributions page. Following September 2021, the option is available under the video description page.

After entering a “search” keyword, a series of Filters become available, as shown in Figure 1. One of these filters, called “Creative Commons”, allows searching for content with a Creative Commons license. The filters are applied on the search query, thus allowing for searching language-specific terms and then have the results filtered with the additional condition of being released under a Creative Commons license.

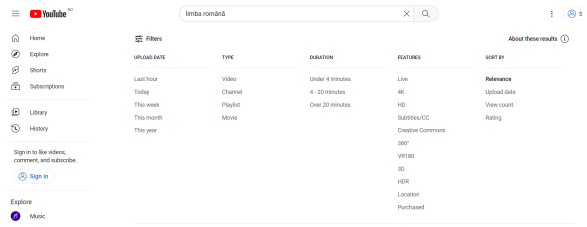

Figure 1: Search filters available in YouTube

A video content released under a Creative Commons license, contains this information in the “Description” section. In the case of long descriptions, the “Show more” button needs to be pressed. This is shown in Figure 2. Once the “Show more” button is clicked, the expanded video description ends with the explicit mentioning of the license, as shown in Figure 3.

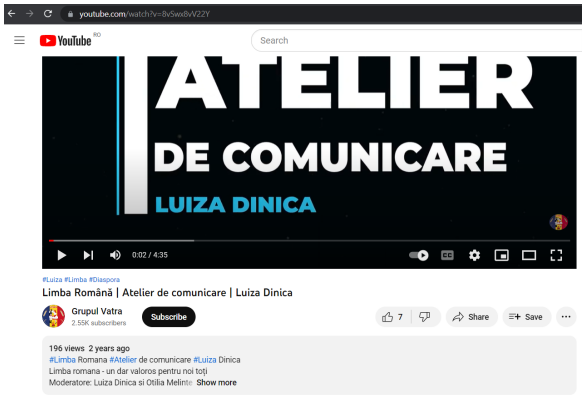

Figure 2: Video with long description

Currently the search interface does now allow a direct search based on the language of the content. Instead search phrases making use of language-specific keywords must be used. Furthermore, it is not possible to search or filter by speaker characteristics. The speaker-related information must be inferred from the description, profile or video content.

[4] https://support.google.com/youtube/answer/2797468?hl=en

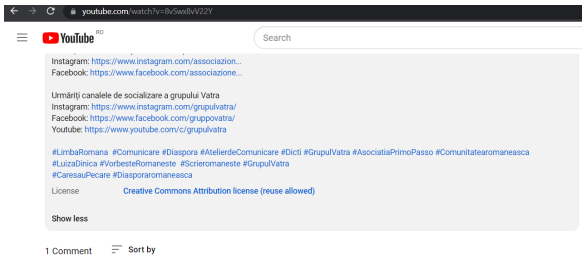

Figure 3: Expanded video content description with the license clearly shown

2.2. Vimeo

Vimeo is another very popular American platform for sharing videos. They boast higher quality of their content, whose important characteristic is being artistic, having high-definition videos. Another important aspect is that the videos have no ads. Users can upload videos with a limit depending on their plan (free or paid).

The "Copyright"[5] page makes it clear that users must upload only materials that "do not infringe any third-party copyright". Thus, users are responsible for the content they provide. The licenses under which videos are released in Vimeo are of the type Creative Commons, 6 such types being in use at the moment of this writing: CC BY (Attribution), CC BY-NC-ND (Attribution-NonCommercial-NoDerivs), CC BY-NC (Attribution-NonCommercial), CC BY-NC-SA (Attribution-NonCommercial-ShareAlike), CC BY-ND (Attribution-NonDerivs), CC BY-SA (Attribution-ShareAlike), as well as CC0 (No Rights Reserved).

The possibility of filtering results becomes available after a keyword is introduced in the search box. Filters help refine searches by the categories to which videos belong (e.g., documentary, animation, art, industry, sports, etc.), High dynamic range (HDR), resolution, duration, price, license, etc. Information about the license of a video is displayed when the "More" button is clicked on the video's page (Figure 4).

The search is possible for different keywords, but not for the language of the video, nor for the speaker's characteristics (age, gender, origin, native language, etc.), which is clearly an obstacle when trying to harvest data according to the criteria of the project's interest.

Other problems encountered with Vimeo while searching for data of interest include: a little number of results with permissive licenses, many results with an unspecified license, videos of long duration, which would unbalance the data if used.

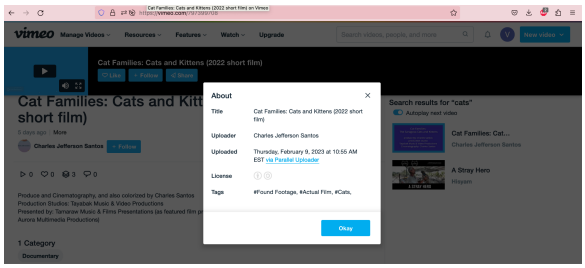

Figure 4: More information about a video, including the license type

[5] https://vimeo.com/dmca

2.3. TikTok

TikTok is a very popular Chinese platform for video hosting, available outside China. The videos here are short, with a maximum duration of 10 minutes. They have music added in the background.

According to the Terms of Service[6] and the Community Guidelines[7], users are not allowed to upload content that infringes a third-party copyright. There is no license to be selected, all uploaded materials being covered by the TikTok license. With regard to parties other than the Platform or its Affiliates, the TikTok Terms of Service mention that "You also grant to each user of the Platform a non-exclusive, royalty-free, worldwide license to access and use your content, including to reproduce (e.g. to copy, share or download), adapt or make derivative works (e.g. to include your content in their content) perform and communicate that content to the public (e.g. to display it) using the features and functions of the Platform for entertainment purposes, subject to your Platform settings." These terms are not entirely clear. On one hand the terms mention the right to reproduce, including "download", on the other hand it mentions "using the features and functions of the Platform for entertainment purposes".

Search results cannot be filtered in TikTok, but searches can be made by users, videos, sounds, and hashtags: in Figure 5 we show how searching for a certain key expression (*sfaturi cosmetice* En. cosmetics tips) can be made in available accounts.

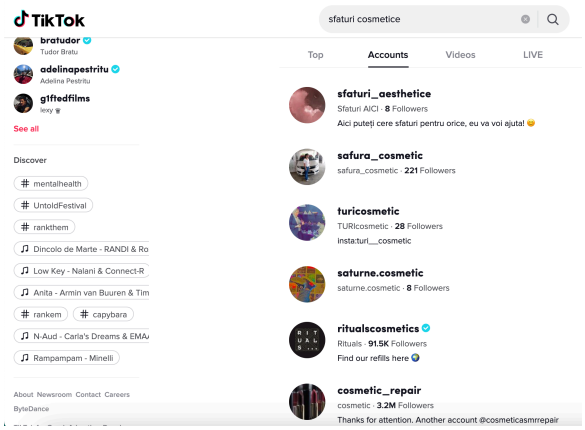

Figure 5: SoundCloud license settings page

2.4. SoundCloud

SoundCloud is a web platform that allows users to host audio recordings of themselves performing music, reciting poems, reading literature, presenting radio shows, etc. The default settings of the account do not allow the content to be freely redistributed and used, and the users have to explicitly tick the Creative Common license for their content when they want to make it freely available (see Figure 6 for the referred to settings page).

SoundCloud has a well-designed search interface which allows for keyword and key-phrase searches, hashtags searches and different types of object types searches (audio clips, users, albums and playlists). Unfortunately, this interface does not allow users to:

- Search clips in a given language (e.g. Romanian).

[6] https://www.tiktok.com/legal/page/eea/terms-of-service/en
[7] https://www.tiktok.com/community-guidelines

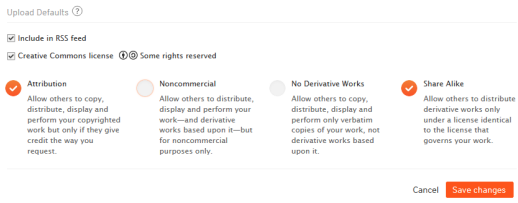

Figure 6: SoundCloud license settings page

- Search clips by usage license (e.g. Creative Commons vs. restricted).

These two drawbacks make SoundCloud a more difficult platform to use to create an audio corpus for a given language. That being said, one can still use SoundCloud to mine for freely available audio clips by executing the following procedure:

- Search for relevant hashtags like #Storytelling sau #Audiobooks or search for specific Romanian phrases such as “povești pentru copii” (stories for kids), “emisiuni radio” (radio shows) or podcasts.
- Carefully check each result obtained in the previous step to ensure that the content is in Romanian, is of reasonable quality and bears the Creative Commons license for the targeted tracks.
- If some track is found to be of interest and possesses the above-mentioned properties, the ”Related tracks” section appearing in the right column of the track page usually points to similar tracks, both in terms of the spoken language and license specification.

For example, following the above steps, we were able to find the account of a radio station that set the Creative Commons license for all its content in SoundCloud (hundreds of hours of aired shows, see Figure 7 for one of their radio tracks marked with the CC license).

Figure 7: One radio station account in SoundCloud with the Creative Commons license set for its content

SoundCloud contains audio tracks of very good quality. The recorded speech is very good, and, with the exception of radio shows which may contain a faint musical background, free

of noise. Searching for tracks for our speech corpus (i.e. in Romanian), we found that SoundCloud mainly hosts the following types of content:

- story reading
- podcasts (e.g. medical, psychology, literature, motivational etc.)
- original music and music mixes
- recordings of religious proceedings
- radio shows
- documentary recordings (e.g. of old people talking about their young lives)

The Romanian content is mostly recorded by either young or middle-aged people, with no obvious gender predominance. The usable tracks are usually long (over an hour, with the exception of short stories).

2.5. LinguaLibre

According to the "About" link of the LinguaLibre website, Lingua Libre is "a project of the association Wikimédia France which aims to build a collaborative, multilingual, audiovisual corpus under free license". The website is created so that anyone can record their voice and store the files using metadata such as the spoken language or the gender of the speaker.

The search interface is very detailed and allows for filters such as spoken language, gender of the speaker and the language proficiency. An example search for Romanian is presented in Figure 8.

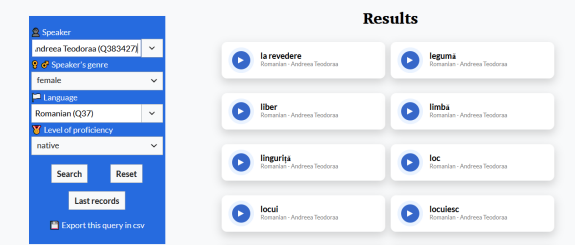

Figure 8: The search results for the LinguaLibre website

The Romanian language is not very well represented in Lingua Libre at the time of writing. There is a rich corpus of spoken single words, from various male and female speakers, but very few phrases and no sentences. This portal is to be monitored for new data that could be used for speech recognition and synthesis tools.

3. Activity 2. Gather a sample of open multimedia content

Using the search features identified in Activity 1, a sample of multimedia content, targeting underrepresented Romanian language speakers was downloaded. For each downloaded material, a record was kept with regard to the source platform, URL, license, date of download and a screenshot of the content as appearing in the platform at the download date. This provision enables us to have a record of the license associated with the material at the date it was downloaded.

The content was identified by using a combination of platform filters (especially for license selection) and custom search keyphrases. The search expressions aimed to identify Romanian speakers (the expressions were formulated in the Romanian language) and interesting content for the project's purposes.

After identification, the content was downloaded as .mp4 video files with a low video resolution (to reduce space requirements). The actual download was performed using an online downloader application[8]. For the project's use, the video content was then placed into a shared Google Drive, available to the project's team.

Table 1 shows a breakdown by platform of the downloaded sample files. Table 2 shows a breakdown by license type.

Platform	Files
YouTube	30
Vimeo	4
SoundCloud	5

Table 1: Content breakdown by platform

License	Files
CC Attribution	32
CC Attribution Non-Commercial	2
CC Attribution Non-Commercial Share Alike	5

Table 2: Content breakdown by platform

During the download process, a screenshot of the video file as it appears in the platform was saved. This allows us to confirm the license under which the content was released at the download date. This screenshot will not be released as part of the corpus since it may contain potentially sensitive information (not available under an open license, or containing personal data) in the form of comments or recommended videos. However, the screenshots will be kept after the project's end date to be able to respond to potential claims that the content was not available under an open license.

A Google Sheet was created to allow gathering metadata information about each multimedia file. This is described in the Annotation Guide, available in Appendix A. As part of this activity, general file-based metadata was collected, such as: annotator, platform, URL, duration, license, speech type (read or spontaneous), quality, and speaker-related information: gender and age. This allowed us to confirm that the collected sample corresponds to the objectives of the USPDATRO project. The Google Sheet was setup in such a way that a total duration was computed and the various values were selected from a drop down list with nomenclature values (platform, license, age, sex, speech type, quality). This allowed for computing statistics on the resulting corpus.

4. Activity 3. Manual annotation of retrieved samples

Files downloaded as part of Activity 2 were manually annotated with subtitles, an indication of the speech usability (noise, superimposed voices, unclear voices, etc.), and type of voice (age interval, gender, license, considering only anonymized characteristics). Part of

[8] https://en.savefrom.net/1-youtube-video-downloader-463/

this information was already gathered in Activity 2, when the samples were downloaded. However, during this activity the information was confirmed and missing data was added.

The actual transcription was performed using the Subtitle Edit application, as described in the Annotation Guide, Appendix A. As indicated in the guide, the transcriptions were saved in CSV format. This was considered the easiest to use with processing scripts. However, it seems in certain cases Subtitle Edit is not able to re-open a CSV file previously saved. To account for this issue, we created a simple script that allows converting CSV files to SRT files, which can then be re-opened. The script is available in the project's GitHub repository[9].

Regarding the transcription of the speech, we encountered the following challenges:

- English words in Romanian speech, including websites, prepositions or adjectives that fit well in the Romanian phrase, e.g. "Aș fi făcut ceva, **like** să merg unde trebuia." (I would have done something, **like** to go where I was supposed to).
- Overlapping voices in the spontaneous recordings; according to the case, we either 1. did a very refined segmentation to separate the voices; 2. left untranscribed short fragments of recordings where the overlap was occurring; 3. ignored longer fragments of recording, even if they contained non-overlapping parts, if they represented clusters of overlapping regions.
- Discerning exactly what sounds the speaker pronounced, especially in spontaneous speech; in many situations, reducing the play rate to 40-50% was necessary to identify the uttered phonemes in words that were, most of the time, recognisable even with the missing sounds.
- Although we aimed for segmentation at sentence boundary as much as possible, for some speakers this was very difficult because of their tendency to systematically both pause in the middle of the sentence and not pause at the end of the sentence.
- Although most of the time the waveform available in SubtitleEdit was very helpful in the segmentation process, sometimes, when very loud sounds where present somewhere in the recording, the relative rendering of the sound intensity for the spoken parts resulted in a flattening of the waveform, thus making it less useful for segmentation.

Statistics from the annotation process are given in Tables 3 and 4. The total duration of the downloaded files is 7h46m18s (as indicated by the originating platform). However, the SoundCloud platform offered especially large files (1h in duration). In order to obtain a dataset with a better distribution across age and gender we decided in the case of large files to consider only partial content. This lead to a total duration of 5h23m21s that was considered for transcription. With regard to the quality of the speech, assessed using a mean opinion score (MOS), we wanted to have good quality sound, thus aiming for MOS=5 or MOS=4. However, we considered also 3 examples with MOS=3 in order to make the dataset useful for testing ASR systems under more difficult conditions.

We tried to detect the age of the speakers as close as possible. In some cases the age was mentioned in the recording, while in some other cases we were able to identify the person in other websites or social media. In a few cases it was not possible to actually find an exact number for the speaker's age. In this case we were forced to make a determination based on common sense (for example a child in a kindergarten was obviously in the "under 14" category). When the speaker's age was determined from additional websites (for example, for some well-known people we were able to identify the birth date) we also took into consideration the moment when the video was released and did the math. For example, a person

[9] https://github.com/racai-ai/USPDATRO

Indicator	Category	# Files
Type of speech	Read	11
	Spontaneous	28
Number of speakers	1	20
	2	13
	3	2
	5	1
	9	3
MOS quality	5	22
	4	14
	3	3

Table 3: File based statistics

having presently 72 years appeared in a video from 8 years ago, thus leading to the conclusion that he speaker was about 64 years when the video was recorded. Thus, we tried to map the speakers as closely as possible to the appropriate age category. The exact age was not recorded, since we considered this to be personal information. Instead only one of the 6 age categories was kept as part of the recorded metadata. In the 19-29 age category we focused on selecting mostly feminine voices. When the speaker is a male in this category, the person has most of the time a supportive role in the recording and has generally few and short interventions: see Table 6 for the total time coverage of the 19-29 male voice category (11m34s) compared to the 19-29 female voice category (54m).

Indicator	Category	# Speakers
Age	< 14	18
	14 - 19	3
	19 - 29	25
	30 - 50	20
	50 - 70	12
	> 70	6
Gender	F	51
	M	33

Table 4: Speaker statistics

Half of the files in the dataset consist of a single speaker. In this case, the recording is usually done by the speaker itself. Files that contain 2 speakers are usually recorded TV shows where one of the speakers is the moderator and the other speaker presents his ideas, thus covering more time. Children (age category "< 14") usually appear in videos with multiple speakers. In this case, we found shows and interviews with kindergarten personnel that also included the children's views on projects or improvements to the kindergarten.

The dataset covers both read (11 files) and spontaneous speech (28 files). In some cases, the speaker was seen reading from a paper or other type of material. In other cases, the speaker was clearly being interviewed on the street or in some other place, or the speaker didn't know how to read (very young kindergarten children). However, in certain cases the exact speech type was not obvious. In this case we used our judgement, considering for example TV news as being read speech (assuming the presenter was reading from a teleprompter

device).

5. Activity 4. Report and dataset release

This activity covered the production of the present report as well as release of the constructed dataset. The experience gathered from the previous activities was summarized in the report. Based on the annotations, the text-aligned voice segments were extracted and released in a dataset usable for evaluation or training ASR systems for Romanian language.

The overall duration of the resulting dataset (the sum of all the extracted segments) is 4h18m55s. Statistics are given in Table 5 and a speaker breakdown by considering both gender and age is given in Table 6.

Indicator	Category	Duration	# Segments
Gender	F	2h8m42s	1,506
	M	2h10m13s	1,131
Age	< 14	10m45s	175
	14 - 19	15m20s	168
	19 - 29	1h5m34s	676
	30 - 50	45m41s	457
	50 - 70	1h1m22s	674
	> 70	1h0m14s	487
MOS	5	2h20m34s	1,187
	4	1h56m44s	1,435
	3	1m37s	15

Table 5: Dataset statistics

Gender	Age	Duration	# Segments
F	< 14	1m38s	27
F	19 - 29	54m	557
F	30 - 50	33m51s	352
F	50 - 70	26m26s	431
F	> 70	12m48s	139
M	< 14	9m06s	148
M	14 - 19	15m20s	168
M	19 - 29	11m34s	119
M	30 - 50	11m51s	105
M	50 - 70	34m56s	243
M	> 70	47m26s	348

Table 6: Speaker breakdown by gender and age

Audio was extracted from the downloaded video files using the FFMPEG command, available on the Linux operating system. The actual command used is given in Figure 9. For the purposes of this command, *from* represents the starting time of the segment, *to* represents the end time of the segment, *dir* is the folder containing the raw videos, *id* is the id of the video file, *fname* is the file name associated with the output audio segment. The segment file

```
ffmpeg -y -ss $from -to $to -i $dir/$id.mp4 -vn -acodec pcm_s16le
       -ac 1 -ar 16000  ${fname}.wav
```

Figure 9: FFMPEG command for converting video to dataset audio files

name follows the format *videoId_segmentNumber.wav* (for example, considering the video with id *1001*, the following are valid segment files: *1001_1.wav*, *1001_2.wav*).

The characteristics of the resulting audio files are given in Table 7.

Characteristic	Value
Channels	1
Sample Rate	16,000
Precision	16-bit
Encoding	16-bit Signed Integer PCM

Table 7: Characteristics of audio files

Each audio segment has a corresponding text file, with the same file name and the ".txt" extension. These text files represent the transcription of the speech content present in the audio file. The files are UTF-8 encoded, with the appropriate Romanian characters. The files were annotated using UDPipe (Straka and Straková, 2017), integrated in the RELATE platform (Păiș et al., 2020, 2019; Păiș, 2020), using a recent model (Păiș et al., 2021). Statistics on the text part were computed in the RELATE platform and are given in Table 8.

Indicator	Value
Text files	2,637
Size (bytes)	237,747
Sentences	6,652
Tokens	48,530
Unique Tokens	8,221
Unique Lemmas	5,509
Hapax legomena	5,055
UPOS Noun	8,471
UPOS Verb	5,793
UPOS Adj	1,952
UPOS Adv	3,717
UPOS Adp	4,009
UPOS Num	615
UPOS PropN	851

Table 8: Characteristics of the text part of the dataset

The dataset was released on the Zenodo platform[10]. The platform offers long term storage, the possibility of creating new versions (if the need will arise in the future) and Digital Object Identifiers (DOIs) for each individual version and for the dataset itself (all versions). The dataset can be cited in scientific papers as (Păiș et al., 2023) (the citation is provided by Zenodo in bib format). A screenshot from the Zenodo platform with the resource is available in Figure 10.

[10] https://zenodo.org/record/7898233#.ZFSXrXZBy3A

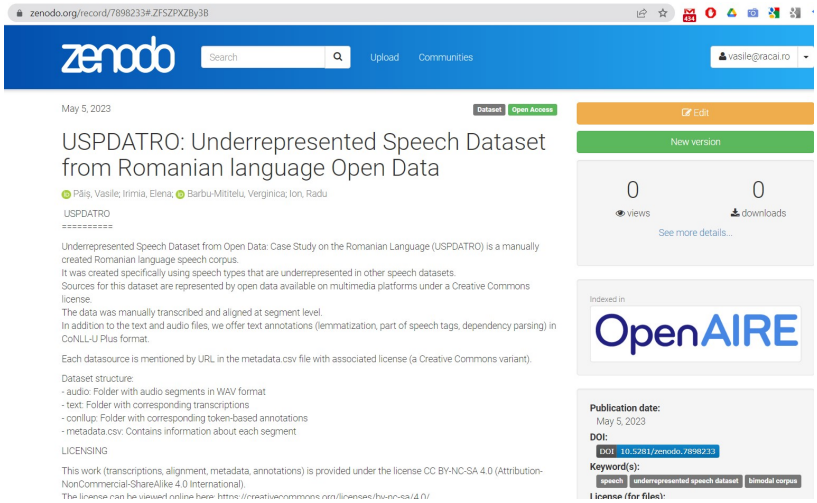

Figure 10: Screenshot from the Zenodo platform with the USPDATRO dataset

The dataset was submitted for publication in the European Language Grid (Rehm et al., 2020) catalogue[11]. Figure 11 shows a screenshot from the ELG catalogue with the USPDATRO corpus.

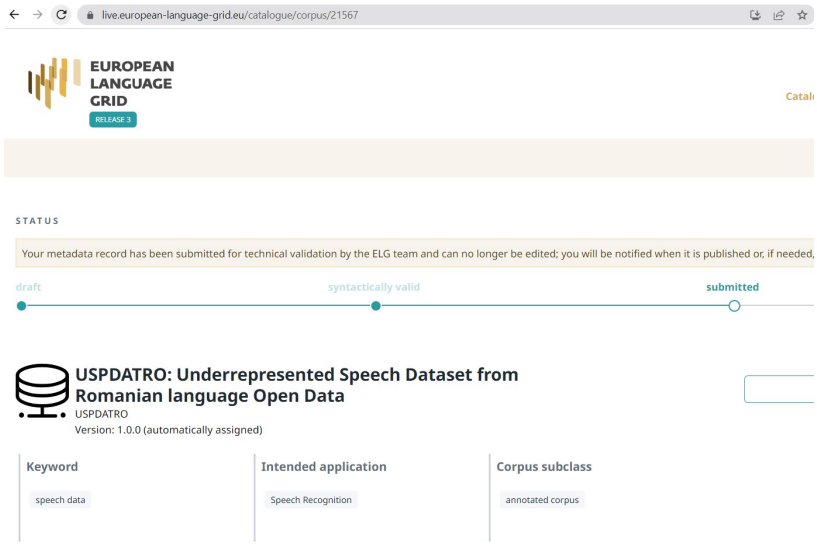

Figure 11: Screenshot from the European Language Grid platform with the USPDATRO dataset

The dataset was further indexed in the RELATE platform dedicated to Romanian language resources and technologies. A screenshot from the platform is available in Figure 12. The RELATE platform offers also a backup download, apart from Zenodo. In addition it also provides a stable URL link[12], without the DOI scheme.

A small website[13] dedicated to the project was created and hosted on our web server to

[11] https://live.european-language-grid.eu/catalogue/corpus/21567
[12] https://relate.racai.ro/repository/uspdatro
[13] https://www.racai.ro/p/uspdatro/

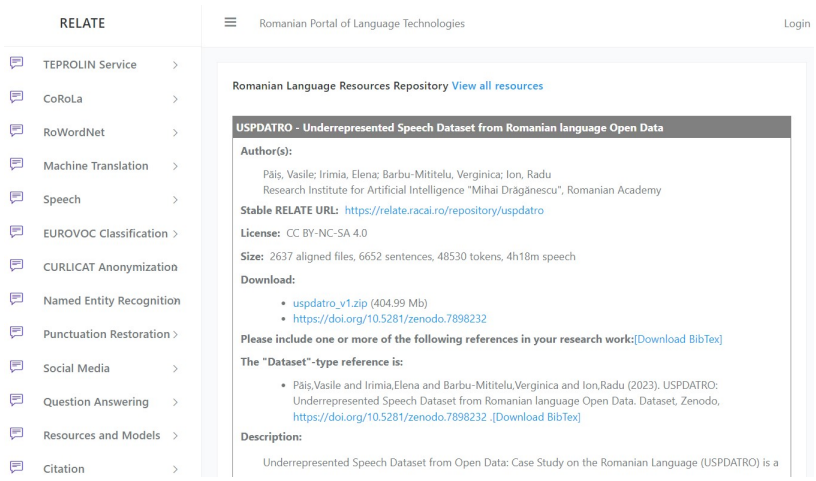

Figure 12: Screenshot from the RELATE platform with the USPDATRO dataset

help with the dissemination of project's results. A screenshot is provided in Figure 13.

Figure 13: Screenshot from the USPDATRO website

6. Summary and Conclusions

The USPDATRO project investigated the usability of open data in mainstream multimedia platforms for building an underrepresented speech dataset. This method allows obtaining training data, useful for speech recognition systems, outside of the commonly available speech types. We focused on obtaining Romanian language speech data, because such resources are small compared to English and other European languages, as asessed during the European Language Equality project (Păiș and Tufiș, 2022). Therefore, besides investigating a new source of speech data, this project aimed to contribute towards increasing the Digital Language Equality Metric (Gaspari et al., 2022) with regard to the Romanian language.

Our findings suggest that multimedia platforms host open content (under Creative Commons or similar licenses) that can be used for research purposes in the domain of building and evaluating ASR systems. The quantity and quality of the content varies from platform to platform. In the case of Romanian language, we found that YouTube hosts larger volumes of

different speech types, compared to other platforms. Nevertheless, we were able to identify usable content in all the investigated platforms.

The methodology applied for the purposes of the USPDATRO project can be extended to other under-resourced languages or speech types. The main challenges to overcome are represented by a lack of "search by language" functionality in the multimedia platforms and the different licenses available. The first challenge is easily overcome by using appropriate search keywords and phrases (examples of Romanian search expressions with associated target groups are available in the Annotation Guide, Appendix A). The different licensing conditions need a lot of attention during the data gathering process. Different Creative Commons licenses allow for different usage types (for example the Non-Commercial variant does not allow for the content to be used for commercial purposes).

Appendix A Annotation guide

A.1 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 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 are given in Table 9.

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:

- 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).
- 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. An example is given in Figure 14.

Keywords	Target group
liceu (En. high-school) elevii te învață (En. the students teach you) sugestii pentru scoala (En. school suggestions) povesti pentru copii (En. children fairy tales) poezie grădiniță (En. poetry kindergarten) lecții online (En. online lessons) părinți și copii (En. parents and children) probleme adolescenți (En. problems teenagers)	Young people
pentru tineri (En. for young people) sfaturi pentru tineri (En. advice for young people) sfaturi duhovnicești (En. spiritual advice) viața la pensie (EN. life at pension)	Older people
emisiune pentru femei (En. women show) feminism și literatură (En. feminism and literature) sfaturi cosmetice (cosmetics tips)	Women
emisiuni radio (En. radio broadcasts) podcast romania romania editura (En. publishing house) antropologie (En. anthropology) psihologie (En. psychology)	Generic

Table 9: Proposed search words and phrases

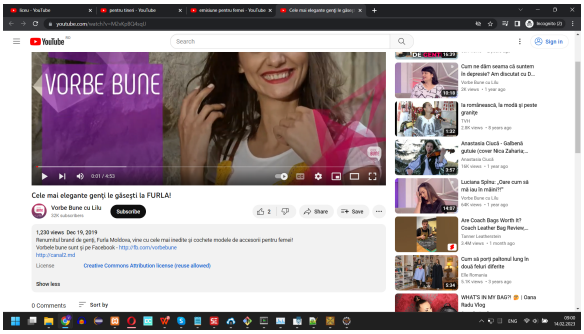

Figure 14: Example screenshot

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 (Figure 15).

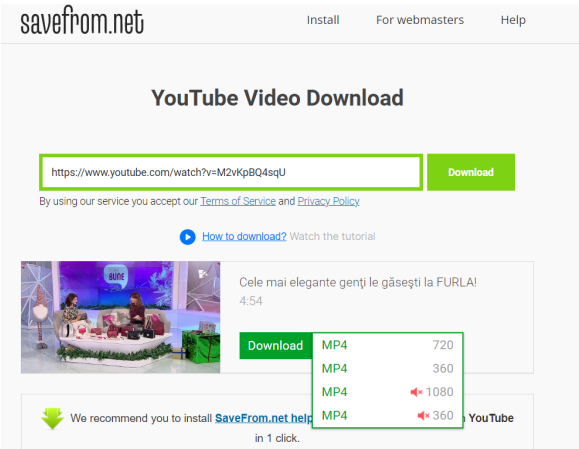

Figure 15: Video download

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.

A.2 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

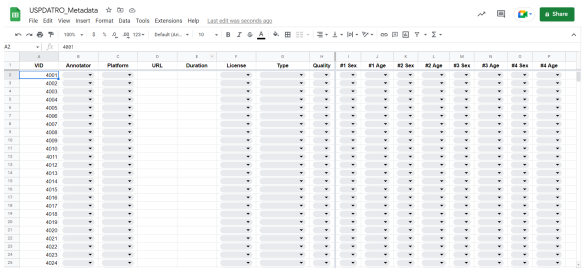

Figure 16: Metadata collection sheet

For storing the metadata a Google Sheets document was setup with data validation rules considering dropdown and allowed data formats (see Figure 16).

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

A.3 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[14], which is a free software under the GNU Public license. Given a new SubtitleEdit project, first the video file is opened (Figure 17).

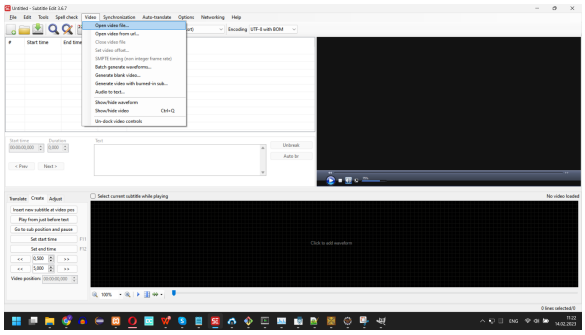

Figure 17: Open video file

The text associated with the audio is entered in the center part of the screen (Figure 18). 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" (Figure 19).

With a click on lower part of the screen the waveform can be visualized in order to allow more precise alignments (Figure 20).

For more efficiency in the transcription work, useful shortcuts can be added and default shortcuts of application can be edited and changed by accessing the Options/Settings menu.

[14] https://www.nikse.dk/, https://github.com/SubtitleEdit/subtitleedit

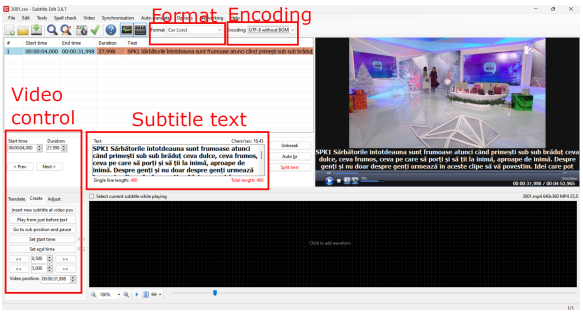

Figure 18: SubTitle Edit main window regions

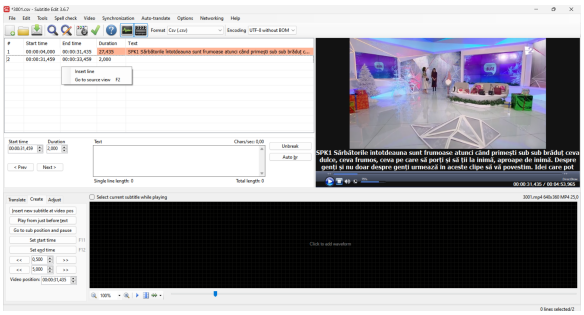

Figure 19: Adding a new text fragment

For example, using Up and Down arrows to set the start/end positions of the text proved to be more efficient that using the mouse or the default F-key shortcuts. Alt-Q was also a good choice for generating a new subtitle.

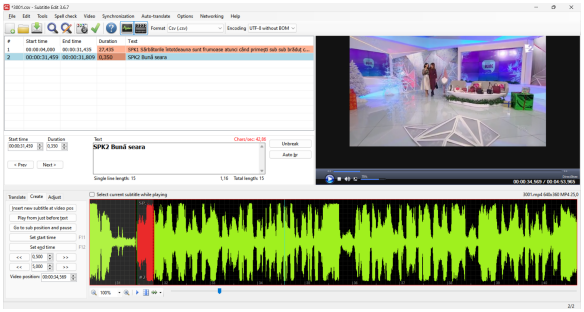

Figure 20: Audio wave visualization

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

A.4 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.

An example is given in Figure 21.

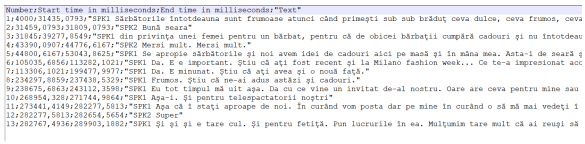

Figure 21: Example CSV file

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

Appendix B Dataset

VID	Platform	URL	Time	Trans.	Lic	Type	MOS
1001	Youtube	https://www.youtube.com /watch?v=IXZRA2jhfGY	0:09:10	0:09:10	CC BY	Read	4
1002	Youtube	https://www.youtube.com /watch?v=k4ve7JnFstg	0:07:32	0:07:32	CC BY	Spont	5
1003	Youtube	https://www.youtube.com /watch?v=KogQQ0H5hNM	0:03:50	0:03:50	CC BY	Spont	5
1004	Youtube	https://www.youtube.com /watch?v=Qtx9rSjrKPw	0:11:16	0:11:16	CC BY	Spont	4
1005	Youtube	https://www.youtube.com /watch?v=apMKXxPOfIQ	0:13:56	0:13:56	CC BY	Spont	4
1006	Youtube	https://www.youtube.com /watch?v=TV-JaJUHsKs	0:14:47	0:14:47	CC BY	Spont	4
1007	Youtube	https://www.youtube.com /watch?v=H1xvuSRPbLw	0:05:17	0:05:17	CC BY	Read	5
1008	Youtube	https://www.youtube.com /watch?v=CWGLM30HI-A	0:10:30	0:10:30	CC BY	Spont	4
1009	Youtube	https://www.youtube.com /watch?v=ePeqwXTumD0	0:19:28	0:19:28	CC BY	Spont	5
1010	Youtube	https://www.youtube.com /watch?v=g_mmLitM6qg	0:10:28	0:10:28	CC BY	Spont	4
1011	Youtube	https://www.youtube.com /watch?v=8ZaGLUcf-C8	0:16:52	0:16:52	CC BY	Spont	5
1012	Youtube	https://www.youtube.com /watch?v=uEwtjDSc9og	0:04:06	0:04:06	CC BY	Spont	5
1013	Youtube	https://www.youtube.com /watch?v=8J8ggtQWVuI	0:17:13	0:04:06	CC BY	Spont	4

2001	Sound Cloud	https://soundcloud.com /urbankid-853010567 /loredana-neagu-mananca-ti-legumele-balauco-povesti-citite-de-parinti	0:06:32	0:06:32	CC BY NC SA	Read	5
2002	Sound Cloud	https://soundcloud.com /user-669345969 /oana-imre-sound-of-soul-editia-18-iulie-2020	1:18:32	0:43:51	CC BY NC SA	Spont	4
2003	Sound Cloud	https://soundcloud.com /urbankid-853010567 /razvan-zlavog-soricel-de-biblioteca-povesti-citite-de-parinti	0:04:25	0:04:25	CC BY NC SA	Read	5
2004	Sound Cloud	https://soundcloud.com /andreirosca /aliniat-cu-tine-povesti-de-viata-erika-popliceanu	1:11:24	0:11:30	CC BY NC SA	Spont	4
2005	Sound Cloud	https://soundcloud.com /postoperator /7-acute-care-surgery	0:39:35	0:04:20	CC BY NC SA	Spont	5
3001	Youtube	https://www.youtube.com /watch?v=M2vKpBQ4sqU	0:04:53	0:04:53	CC BY	Spont	5
3002	Youtube	https://www.youtube.com /watch?v=e9GgH8qI420	0:01:03	0:01:03	CC BY	Read	5
3003	Youtube	https://www.youtube.com /watch?v=jGqkm27jcLQ	0:08:36	0:08:36	CC BY	Spont	5
3004	Youtube	https://www.youtube.com /watch?v=seNnOWM3Er4	0:06:01	0:06:01	CC BY	Spont	5
3005	Youtube	https://www.youtube.com /watch?v=R_7IRUypsIE	0:11:06	0:11:06	CC BY	Spont	4
3006	Youtube	https://www.youtube.com /watch?v=r24br-GK45I	0:18:14	0:18:14	CC BY	Spont	5
3007	Youtube	https://www.youtube.com /watch?v=0TkGoII_jQQ	0:01:56	0:01:56	CC BY	Spont	3
3008	Youtube	https://www.youtube.com /watch?v=m6jsVpvcBaI	0:02:25	0:02:25	CC BY	Spont	5
3009	Youtube	https://www.youtube.com /watch?v=La4WLyVzXy0	0:13:56	0:13:56	CC BY	Spont	4
3010	Youtube	https://www.youtube.com /watch?v=rAz8KZQP_FQ	0:05:08	0:05:08	CC BY	Read	5
3011	Youtube	https://www.youtube.com /watch?v=bukJfp_yzm0	0:00:56	0:00:56	CC BY	Spont	3
3012	Youtube	https://www.youtube.com /watch?v=CGxllcjdDHg	0:00:14	0:00:14	CC BY	Spont	4
3013	Youtube	https://www.youtube.com /watch?v=VbxytH8veCY	0:03:31	0:03:31	CC BY	Read	4
3014	Youtube	https://www.youtube.com /watch?v=–4WT21kAYc	0:02:04	0:02:04	CC BY	Spont	5
3015	Youtube	https://www.youtube.com /watch?v=tVGgTyIzlsU	0:01:15	0:01:15	CC BY	Read	4

3016	Youtube	https://www.youtube.com /watch?v=q0UwZ1tCTBE	0:12:11	0:12:11	CC BY	Spont	5
3017	Youtube	https://www.youtube.com /watch?v=AqPrZs63u6c	0:10:38	0:10:38	CC BY	Spont	5
4001	Vimeo	https://vimeo.com /133998030	0:01:24	0:01:24	CC BY	Read	5
4002	Vimeo	https://vimeo.com /211469494	0:03:00	0:03:00	CC BY NC	Read	5
4003	Vimeo	https://vimeo.com /157569382	0:02:06	0:02:06	CC BY NC	Read	3
4008	Vimeo	https://vimeo.com /55586045	0:10:48	0:10:48	CC BY	Spont	5

References

Federico Gaspari, Owen Gallagher, Georg Rehm, Maria Giagkou, Stelios Piperidis, Jane Dunne, and Andy Way. Introducing the Digital Language Equality Metric: Technological Factors. In Itziar Aldabe, BegoÃ±a Altuna, Aritz Farwell, and German Rigau, editors, *Proceedings of the Workshop Towards Digital Language Equality (TDLE 2022; co-located with LREC 2022)*, pages 1–12, Marseille, France, 6 2022. 20 June 2022.

Vasile Păiş, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, and Dan Tufiș. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401, 2021. URL https://www.romjist.ro/full-texts/paper700.pdf.

Vasile Păiș. Multiple annotation pipelines inside the relate platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75, 2020. URL https://profs.info.uaic.ro/~consilr/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=73.

Vasile Păiș and Dan Tufiș. Deliverable D1.29 Report on the Romanian Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_29__Language_Report_Romanian_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Vasile Păiș, Dan Tufiș, and Radu Ion. Integration of romanian nlp tools into the relate platform. In *International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 181–192, 2019. URL https://profs.info.uaic.ro/~consilr/2019/wp-content/uploads/2020/01/ConsILR2019_final_BTT-60-ex-B5.pdf#page=189.

Vasile Păiș, Radu Ion, and Dan Tufiș. A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France, 2020. European Language Resources Association. ISBN 979-10-95546-64-1. URL https://www.aclweb.org/anthology/2020.iwltp-1.13.

Vasile Păiș, Elena Irimia, Verginica Barbu-Mititelu, and Radu Ion. USPDATRO: Underrepresented Speech Dataset from Romanian language Open Data, May 2023. URL https://doi.org/10.5281/zenodo.7898233.

Georg Rehm, Maria Berger, Ela Elsholz, Stefanie Hegele, Florian Kintzel, Katrin Marheinecke, Stelios Piperidis, Miltos Deligiannis, Dimitris Galanis, Katerina Gkirtzou, Penny Labropoulou, Kalina Bontcheva, David Jones, Ian Roberts, Jan Hajic, Jana Hamrlová, Lukáš Kačena, Khalid Choukri, Victoria Arranz, Andrejs Vasiļjevs, Orians Anvari, Andis Lagzdiņš, Jūlija Meļņika, Gerhard Backfried, Erinç Dikici, Miroslav Janosik, Katja Prinz, Christoph Prinz, Severin Stampler, Dorothea Thomas-Aniola, José Manuel Gómez Pérez, Andres Garcia Silva, Christian Berrío, Ulrich Germann, Steve Renals, and Ondrej Klejch. European Language Grid: An Overview. In Nicoletta Calzolari, Frédéric

Béchet, Philippe Blache, Christopher Cieri, Khalid Choukri, Thierry Declerck, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, pages 3359–3373, Marseille, France, 5 2020. European Language Resources Association (ELRA).

Milan Straka and Jana Straková. Tokenizing, pos tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf. DOI: 10.18653/v1/K17-3009.