# CoBiLiRo: A Research Platform for Bimodal Corpora

**Dan Cristea[1,2], Ionuț Pistol[1], Șerban Boghiu[1], Anca-Diana Bibiri[3], Daniela Gîfu[1,2], Andrei Scutelnicu[1,2], Mihaela Onofrei[1,2], Diana Trandabăț[1], George Bugeag[1]**

[1]"Alexandru Ioan Cuza" University of Iași, Faculty of Computer Science (UAIC-FII)
[2]Institute of Computer Science, Romanian Academy, Iași Branch (ARFI-IIT)
[3]"Alexandru Ioan Cuza" University of Iași, Institute for Interdisciplinary Research, Social Sciences and Humanities Research Department (UAIC-ICI)
[1]16, Berthelot St., 700483 Iași, Romania,
[2] Dr. Theodor Codrescu St., 700481 Iași, Romania,
[3]54, Lascăr Catargi St., 700107, Iași, Romania

{danu.cristea, pistol.ionutcristian, serbanboghiu, anca.bibiri, daniela.gifu73, andreiscutelnicu, mihaela.plamada.onofrei, diana.trandabat, bugeag.george}@gmail.com

**Abstract**

This paper describes the on-going work carried out within the CoBiLiRo (Bimodal Corpus for Romanian Language) research project, part of ReTeRom (Resources and Technologies for Developing Human-Machine Interfaces in Romanian). Data annotation finds increasing use in speech recognition and synthesis with the goal to support learning processes. In this context, a variety of different annotation systems for application to Speech and Text Processing environments have been presented. Even if many designs for the data annotations workflow have emerged, the process of handling metadata, to manage complex user-defined annotations, is not covered enough. We propose a design of the format aimed to serve as an annotation standard for bimodal resources, which facilitates searching, editing and statistical analysis operations over it. The design and implementation of an infrastructure that houses the resources are also presented. The goal is widening the dissemination of bimodal corpora for research valorisation and use in applications. Also, this study reports on the main operations of the web Platform which hosts the corpus and the automatic conversion flows that brings the submitted files at the format accepted by the Platform.

**Keywords:** bimodal corpus, annotation standard, web platform, speech and text processing, metadata of linguistic resources, CoBiLiRo, ReTeRom.

## 1. Introduction

In this paper we present CoBiLiRo, an environment intended to act as a hosting, editing and processing platform for large collections of parallel speech/text data. In actual use now for the data of the ReTeRom project, CoBiLiRo contains a collection of bimodal files on Romanian language. The researchers in the ReTeRom project belong to four natural language processing laboratories[1] in Romania that work on speech understanding, speech synthesis, text processing, alignment of speech - text resources and organisation of big repositories of language data for research and public use.

With the purpose to support future research on speech and text technologies dedicated to Romanian, we have done a careful inventory of existing bimodal resources at ReTeRom partners places and have acquired new donations from external providers. The Platform harmonizes the representations of these resources, their annotation and metadata formats, the final aim being to organise the existent and future resources and open large access to bimodal corpora for research valorisation and use in applications.

## 2. Similar Achievements

The reasons for keeping records of speech worldwide are very diverse. A brief enumeration should include: preservation of samples of dying languages, preservation of regional varieties of languages (e.g. the International corpus of English[2], which is an electronic corpus of regional varieties of English throughout the world: Great Britain, Ireland, New Zealand, Canada, Singapore, The Philippines), samples of language in evolution for diachronic comparative studies, interviews with famous people – for cultural heritage preservation. Since 2007, ELRA (European Language Resources Association) organises and distributes[3] a huge collection of language resources, in more than 70 languages and language varieties, among which many are speech or bimodal corpora (Mapelli et al., 2018). Famous speech corpora are: Santa Barbara Corpus of Spoken American English[4] (a large body of recordings of naturally occurring spoken interaction from all over the United States), Cambridge International Corpus[5] containing many other written and spoken corpora (Cambridge and Nottingham Corpus of Discourse in English (CANCODE), Cambridge and Nottingham Spoken Business English (CANBEC), Cambridge Cornell Corpus of Spoken North American English), The Buckeye Speech Corpus[6] (conversational speech on different themes), Bavarian Archive for Speech Signals Corpora (Siemens Synthesis Corpus)[7].

[1] "Mihai Drăgănescu" Research Institute for Artificial Intelligence of the Romanian Academy in Bucharest (RACAI), as ReTeRom Coordinator, Technical University of Cluj-Napoca (UTCN), Politehnica University of Bucharest (UPB) and "Alexandru Ioan Cuza" University of Iași (UAIC).

[2] http://ice-corpora.net/ice/index.html
[3] Via its operational body ELDA.
[4] https://www.linguistics.ucsb.edu/research/santa-barbara-corpus
[5] https://www.cambridge.org/elt/corpus/international_corpus.htm
[6] https://buckeyecorpus.osu.edu/
[7] https://www.phonetik.uni-muenchen.de/forschung/Bas/BasKorporaeng.html



*Proceedings of the 1st International Workshop on Language Technology Platforms (IWLTP 2020)*, pages 22–27
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Much less numerous are bimodal speech-text corpora[8] i.e. collections that keep voices and their transcribed text. Examples are: Turkish bimodal corpus (Polat and Oyucu, 2020), GermaParl: Corpus of Plenary Protocols of the German Bundestag (Blaette, 2017), The Spoken Dutch Corpus (representing contemporary standard Dutch as spoken by adults in The Netherlands and Flanders) (Oostdijk, 2000), C-ORAL-ROM – a multilingual corpus of spontaneous speech (around 1.200.000 words) representing four main Romance languages: French, Italian, Portuguese and Spanish (Cresti and Moneglia, 2005).

Platforms offering access to speech and bimodal resources are already available, perhaps the most significant ones due to their size and inclusion of Romanian language documents are the LRE Map and Clarin's VLO. Both of them include multiple resources of similar nature to those for which we designed and built the CoBiLiRo platform. More complex features are offered by the Virtual Language Observatory (VLO), which allows users to input search queries over the available resources using a custom designed syntax. Also, in VLO one can match resources with available processing tools, the interface indicating which of the available processing tools are compatible with the viewed resource. A functionality of this type is not implemented in CoBiLiRo, since our platform is designed specifically for aligned (speech-text) resources and includes special features allowing users to locate, filter and access such corpora. For example, none of the two platforms mentioned above allow users to search only male voices, only resources of a certain size or to process the available resources (annotate or convert them to a different format).

## 3. Architecture and Functionalities

### 3.1 Technologies and Architectural Patterns

ASP.NET Core is a high-performance, cross-platform, open-source framework used to develop the Cobiliro platform. The web application is hosted on premises in the "Alexandru Ioan Cuza" University of Iași, on a CentOS machine.

The Model-View-Controller (MVC) architecture is used in order to separate the platform into three main groups of components that can be easily extended and modified, each one having its own role:

- *Models*

Models are used to represent database tables and relationships between them. In order to manipulate the data, we use Entity Framework – an object relational mapper that provides a fast way of interacting with the databases and models. Model Validation techniques both server-side and client-side were used in order to ensure that the inserted data is consistent and reliable.

- *Views*

In order to provide a user-friendly interface, we decided to use jQuery and Razor syntax, which offers a way of creating server-side dynamic pages that receives and display data from the models.

- *Controllers*

Controllers handle the web application requests. The services are injected into controllers for achieving Inversion of Control between them and their dependencies. The dependency injection pattern implements inversion of control and assures a loosely coupled web application.

### 3.2 Security

The authentication, authorization, and role management of the application was implemented using the Identity Framework. This framework provides a powerful API that allows us to manage access control and security concerns regarding data privacy with respect to GDPR regulations[9]. Every password was hashed using the PBKDF2 algorithm[10] which is considered to be the safest encryption algorithm and also the most widely used by most applications.

In order to authenticate every HTTP request, we have attached a token (also known as a bearer token) to it. This assures that only allowed users access the shared content. ASP.NET framework also offers an easy mechanism that can facilitate protection for SQL Injection or Cross-Site-Request Forgery attacks.

### 3.3 Data Base

For persistent data storage we have used MariaDB, a free, open-source relational database. Pomelo Entity Framework is an Entity Framework provider that allows use of Entity Framework with a MySQL database.

### 3.4 REST API

Representational state transfer (REST) is a software architectural style that defines a set of constraints to be used for creating Web services. For example, for listening to a sound file, the files and their metadata should be sent as a byte array to the client-side application. The requests are going through our authorization and authentication filters. The serialization is done using the Javascript Object Notation (JSON) which is an open-standard file format.

### 3.5 External NLP Services - TEPROLIN

For the processing of texts that are uploaded as part of the bimodal resources, the Teprolin Web Service[11], developed by RACAI partner, is used. This service allows several operations to be applied to texts, such as:

- restoration of diacritics
- phonetic transcription of words
- converting numbers into their text spellings
- bordering into sentences
- tokenization
- POS-tagging
- lemmatization
- Named Entity Recognition
- NP-chunking

[8] Except for the very frequent sound and video, other bimodal corpora include speech and sign language, or sign language and text.

[9] https://gdpr-info.eu/

[10] https://en.wikipedia.org/wiki/PBKDF2

[11] http://89.38.230.23:5000/

- syntactic parsing, etc.

### 3.6 Functionalities

The Portal is opened to the following categories of users: *administrator*, *resource curator* (responsible for the monitoring and management of new resources), *donator* (a user that offers and uploads their resources, and which can do anything they want with their own resources, including deletion), and *ordinary user* (only for consultation, browsing, therefore having a passive role, or interested in doing theoretical or applied research with the Portal's resources). The access of ordinary users is restricted by IPR, each resource being paired with a specific IPR contract. The hierarchy of rights is: administrator > curator > donator > ordinary user.

The resources on the Platform can be interrogated by different criteria, matching keywords against the description field of the metadata and/or combining other different metadata values. Once found, a resource can be: consulted, by browsing its content with web GUIs, downloaded, deleted, upgraded/updated (delete + upload), or converted to a different format (see Section 4). The whole repository is backed-up periodically. Global statistics on the whole collection are automatically updated and can be consulted at the level of an ordinary user. Other functionalities offered by the Portal include: secured administration panel, responsive design (adapted for mobile devices), newsletter, contact forms with in and out email service for external users, forum of discussions and chat, RSS, Google Analytics.

## 4. Data Formats and Convertors

As part of the process of building the CoBiLiRo repository, we have contacted owners of speech/text resources open to the idea of offering them for research tasks. We have identified three types of original formats that pair speech and text components. This variety, well documented as a project delivery (Trandabăț, 2018), is as follows:

- PHS/LAB, a format which separates text, speech and alignment in different files;
- MULTEXT/TEI, a format described initially in the MULTEXT project and later used by various language resource builders;
- TEXTGRID, a format supported by a large community of European developers and used in a large set of existing resources.

The generous research and development goals that we envisage around the use of the CoBiLiRo platform, all shaped for the purpose of functioning as a sharing and distribution host of bimodal resources, imposes the adoption of a standard resting format for all hosted elements. Taken as an internal standard, this format will allow interchangeability of any types of resources and one-time implementation of a large spectrum of searching, editing and statistics functionalities. This format (Cristea et al., 2018) is inspired by the TEI-P5.10 standard (Sperberg-McQueen and Burnard, 2018), while also including elements from other proposals (Li and Yin, 2007). The TEI-P5 standard has been simplified in some respects and augmented in others to best accommodate the requirements of our bimodal corpora of speech and text data. To organise the functionality of the Platform around this standard, input and output converters have been implemented to support in and out transfers. Central to this format, as will be seen below, is the idea of alignment between the speech and the text components.

The platform also includes an API able to automatically detect the format of the original uploaded resource and launch on it the proper convertor, in order to bring the input files to the standard format. The conversion process is performed on our servers without requiring any user input. Once converted to the standard format, the file can benefit from the Platform's search, editing and statistics capabilities. However, the original format is also preserved, at least for the reason that the conversion is not always lossless. At any time, the user has the option to download either the original format or the CoBiLiRo standard variant of a resource, the last one opening the door for enhanced integration with the Platform.

The CoBiLiRo format includes metadata, kept in a header, and content. The header records:

- source of the object stored,
- gender of speakers,
- identity of speakers (when they agreed to be nominated – as, for instance, in public speeches),
- voice's type (spontaneous or voice-in-reading),
- recording conditions (in lab, noisy environments, etc.),
- duration,
- type of speech files (mp3 or wav),
- speech-text alignment level (sentence, word),
- etc.

These pieces of information are stored in appropriate xml tags and attributes, within the teiHeader tag.

In the content part, segmentation of speech and its alignment with the text is marked. The most common levels of segmentation and alignment are the sentence and the lexical tokens. Since, in the voice files, sentences could sometimes be difficult to border, morphological units (such as words) and phonological elements (phonemes) constitute other possible segmentation elements. More higher layers of annotation could be added: on the speech signal – prosodic annotation (pitch, raise and decrease of the fundamental frequency), and on the textual component – sub-syntactic (nominal groups, clauses, etc.) and syntactic (parsing trees), performed with TEPROLIN services, as shown in Section 3.5.

The CoBiLiRo format allows for three types of segmentation and speech-text alignment, marked using `<unit>` tags. The first type, called "`file`", is adequate for resources held in multiple files. A `<unit>` tag includes child nodes: the `<speech>` child names the file containing the speech component and the `<text>` child points to the corresponding textual transcription file.

The second type of segmentation, called "`start-stop`" (see Figure 1), is adequate for resources that include only one speech file, which is segmented and aligned at temporal boundaries, the text being reproduced between each two such consecutive markers, given in seconds, with the `start` and `stop` attributes.

Finally, the third type, called "`file-start-stop`", represents a combination of the two types presented above.

```
<units>
  <unit>
    <speech speechFile="9C6c_86a.wav" />
    <subunit>
      <speech start="0" stop="0.1881085" />
      <text>"</text>
    </subunit>
    <subunit>
      <speech start="0.1881085" stop="0.2871186" />
      <text>\'1a"</text>
    </subunit>
    <subunit>
      <speech start="0.2871186" stop="0.33094275" />
      <text>n"</text>
    </subunit>
    <subunit>
      <speech start="0.33094275" stop="0.5378901" />
      <text>a::"</text>
    </subunit>
    <subunit>
      <speech start="0.5378901" stop="1.10175" />
      <text>"</text>
    </subunit>
  </unit>
```

Figure 1: An example of a start-stop segmentation and alignment marking (the <text> segments are specific characters, correctly decoded by the interface)

## 5. Data Acquisition, IPR, and Distribution

The audio components include television and radio programs, interviews, public speeches (as those delivered in Parliament or in public events), lectures, movies, theatre plays, read literary works, spontaneous short recordings collected on the street, and other types of speech recordings.

In this paper we classify speech recordings following three criteria. The first criterion takes into consideration the recording act:

- *spontaneous speech*, represented by: narrative voices, dialogs, MapTasks (validated technique in which two subjects work together to complete the task of navigating through a map by describing a route) (Bibiri *et al.*, 2012), appointment-tasks and meetings, "Wizard of Oz" simulations (interactions of human beings with computers for modelling real-life situations) (Bernsen *et al.*, 1998);
- *read speech*, as: chapters from books (or entire books, for instance *Mara*, by Ioan Slavici), news broadcasts, lists of words, number sequences, short sentences (as in the case of the RASC corpus[12]), etc.

The second classification criterion takes into consideration the source of the resource:

- acquired or originally recorded during previous national or international projects *for research purposes*;
- *ad-hoc acquisitions*, which are offered from generous contributors.

Finally, the third criterion considers the intention behind the creation of the resource:

- originally created with the purpose to *develop and improve speech technologies* for Romanian language (such as those created by consortium partners RACAI, UPB and UTCN),
- created for *linguistic, phonological and/or dialectal research* (in general, those created at UAIC).

To take one example, *read speech* resources, acquired for *research purpose* related to *dialectical investigations* offer the opportunity to analyse: various pronunciations in different dialects; the pronunciation specific to males and females; flapping across word boundaries in spontaneous speech; the effect of disfluencies on neighbouring words; duration of sounds at the end of an utterance (in accordance with the feelings expressed); the pronunciation of unstressed vowels (especially at the end of the words); sounds deletion; palatalization across word boundaries – Moldavian dialectal pronunciations, like: *g'ine* (for *bine*; EN: *good*); *k'atră* (for *piatră*; EN: *stone*), *hier* (for *fier*; EN: *iron*); or intonational patterns characterizing Romanian language.

At the moment of writing this paper, the following resources are hosted by the CoBiLiRo Platform. According to the above mentioned criteria, in the category of spontaneous speech corpora there are included: the CoRoLa[13] corpus, *the Reference Corpus for Contemporary Romanian* – supplied by ARFI-IIT and RACAI; the IIT corpus, containing radio debates and interviews – contributed by ARFI-IIT; the SoRoEs corpus, acquired in the project *Romanian and Spanish contrastive intonation analysis. A sociolinguistic approach* – contributed by UAIC-ICI; the Spontaneous Speech Corpus (SSC-train), Spontaneous Speech Corpus (SSC-eval) and Spontaneous Speech Corpus 2 (SSC-eval2) – all contributed by UPB. For read speech corpora, the following resources are uploaded: the Read Speech Corpus, including TV news and talk-shows – contributed by UPB; SWARA (*Mobile System for Rehabilitative Vocal Assistancee of Surgical Aphonia*); a large expressive Romanian speech corpus, reproducing the novel *Mara* written by Ioan Slavici in 1906, in an audiobook format, and Ro-GRID, short recordings with a fix format – all provided by UTCN. The lastly acquired resources consist of 74 hours or recordings, radio interviews, therefore spontaneous speech, ad-hoc acquisitions, offered to improve speech technologies: the "100 Years of Romania" corpus[14] , the "Guess Who's Coming to Dinner" corpus[15], and the "Conversations on culture and science" corpus[16]. All resources are bimodal, therefore including both audio files and transcripts, and the speech-to-text alignments are now being generated by the TADARAV[17] aligner (Georgescu et. *al.*, 2019). In total, the Portal includes now more than 520 hours of speech recordings and their transcriptions.

For all these resources we have agreed and signed with the donor's specific formulations of IPRs, which state also

[12] http://rasc.racai.ro/; https://speech.utcluj.ro/swarasc/

[13] http://corola.racai.ro/

[14] Contributed by prof. Gheorghe Iacob, "Alexandru Ioan Cuza" University of Iași.

[15] Contributed by Vasile Arhire, Romanian Public Television (TVR) Iași.

[16] Contributed by prof. Eugen Munteanu, in conversation with acad. Viorel Barbu.

[17] Same as CoBiLiRo, TADARAV is a component part of ReTeRom.

the distribution rules. The Platform offers more types of access: only consultation of titles (open to any user), access to samples of files, restricted and unrestricted download.

## 6. CoBiLiRo as a Source of Applications and Student Work

Since the main purpose of building the CoBiLiRo Platform was to facilitate research and development of processing tools for Romanian spoken and written language, we already envisioned a list of projects, addressed to UAIC students in Computer Science that would make use of the resources and tools hosted by the Platform. Passed to a class of bachelor 3rd year students enrolled in the course *Techniques of Human Language Engineering* and to the master students in Computational Linguistics, some of these ideas are currently under design and development. We present few of them below.

"*The speaking dictionary*" refers to enhancing an electronic dictionary of the Romanian language, the Thesaurus Dictionary of the Romanian Language in Electronic Format - eDTLR (Cristea et *al.*, 2011; Pătrașcu et *al.*, 2016), with pronunciations for its entry words, as they have been discovered in the bimodal resources hosted by the Platform. This will be accomplished following these steps:

1. All text components of bimodal corpora hosted by CoBiLiRo are lemmatised. Lemmatization (same as POS-tagging) follows the conversion process described in Section 4 of this paper.
2. Resources are aligned at word level between the speech and the text components using the TADARAV aligner, and each alignment is accompanied by an estimated accuracy[18].
3. For each dictionary entry we look for morphologically flexed forms of this lemma in the aligned textual documents. For many dictionary entries it is normal to find more matches. In this case, the specifications of the project require that the candidates be listed and uttered, in the descending order of their estimated accuracy, as long as the user keeps pushing a "Pronunciation" button.

Other project ideas involve adding speech components to previously developed applications. Here follow brief descriptions of three of them.

In "*My speaking diary*", the user can interact with an automatically built diary, by asking questions and listening to answers about the activities she/he has been involved in during the day. By using the device's GPS, the API running on the mobile device can log down the list of pairs <time, place> for the places the user has been located at all along the day. Then, by using a GIS[19], it can associate names to these locations and, using calendar entries and/or an ontology of locations, associate typical activities to these locations. Then, the application can use this information to answer questions such as "How much time did I spent at work today?", "How many times did I go shopping last week?" etc. The generation of spoken answers will be done by using the CoBiLiRo bimodal repository and the technologies developed as part of the associated SINTERO project[20] (Stan and Giurgiu, 2018). Used by Alzheimer patients in incipient phases, the application can delay the boost of the illness.

The project "*I dialogue with the book I read*" will implement an idea uttered in a previous lab project (Cristea et al., 2015), in which we showed how semantic relations between characters of a book can be deciphered in a text. But, vocally interacting with a book content could be extremely attractive for a passionate or a young reader. In this project we want to allow a user that reads a novel from the screen of a device to ask an electronic assistant to bring her/him back to the page where, for instance, Vinicius met Ligia for the first time (from H. Sienkiewicz: *Quo Vadis*), or where Adam loses his father, as he is imprisoned by the police (from Tash Aw: *The map of the invisible world*), or where the kinship relationship between two characters has been explicitly uttered (were there are too many, as in *Forsyte Saga* of John Galsworthy).

As the GPS of the user's mobile seizes the instantaneous location where she/he is located while walking through a city, the application "*Reading while walking*" utters in the user's earphones passages of literature that mention that street, park or another place the user actually happens to be. Thus, traveling in a city is complemented with an enjoyable literary experience.

## 7. Conclusions

CoBiLiRo, a very young accomplishment of the ReTeRom complex project, is a platform that aims to create a repository containing a vast collection of synchronised audio and textual resources, annotated on different levels on both the acoustic and the linguistic components. It will soon become the most significant speech & text repository for the Romanian language, addressing future developments of human-machine interfacing technologies.

After making a careful inventory of existing bimodal resources at partners, we continued to procure more and upload them on the Portal. Meanwhile, our partners in the ReTeRom project already use the material acquired there for speech-text alignment in view of further audio and linguistic experiments, out of which training speech-to-text and text-to-speech processes represent the principal objectives. Tools to harmonize the representation, the annotation and the metadata formats of all these resources are hosted on the Platform. It accounts also for a wide dissemination of the Romanian bimodal corpora, in benefit of research valorisation and usage in applications.

## 8. Acknowledgements

This work was supported by a grant of the Ministry of Research and Innovation, Program PN-III-P1-1.2.-PCCDI, nr. 73/2018, as part of the ReTeRom project.

[18] Feature under development.

[19] Geographical Information System

[20] https://speech.utcluj.ro/sintero/, also a component part of ReTeRom.

## 9. Bibliographical References


Bernsen, N. O., Dybkær, L. and Dybkær, H. (1998). Wizard of Oz Simulation. In *Designing Interactive Speech Systems.* Springer, London, https:doi.org/10.1007/978-1-4471-08979_5, ISBN:978-3-540-76048-1.

Bibiri, A.-D., Turculeț, A. and Panaite Beldianu, O. (2012). The use of the MapTask technique in the projects: *Atlas Multimedia Prosodique de l'Espace Roman (AMPER-ROM)* and *Atlasul Multimedia Prozodic Român (AMPRom)*. In Iulian Boldea (Ed.), *Communication, context, interdisciplinarity. Studies and articles* – in Romanian, vol. II, Publishing house of „Petru Maior" University, Târgu-Mureș, 982-993, ISSN: 2069-3389.

Blaette, A. (2017). *GermaParl. Corpus of Plenary Protocols of the German Bundestag*. TEI files, https://github.com/PolMine/GermaParlTEI.

Cresti, E. and Moneglia, M. (Eds.). (2005). *C-ORAL-ROM: Integrated Reference Corpora for Spoken Romance Languages*, John Benjamins, Amsterdam/Philadelphia.

Koržinek, D., Marasek, K. Brocki, Ł. and Wołk, K. (2017). Polish read speech corpus for speech tools and services. *Selected papers from the CLARIN Annual Conference 2016*. Linköping Electronic Conference Proceedings 136: 54–62.

Cristea, D., Gîfu, D., Colhon, M., Diac, P., Bibiri, A-D., Mărănduc, C., and Scutelnicu, A. (2015) Quo Vadis: A Corpus of Entities and Relations. In: Language, Production, Cognition, and the Lexicon. Text, Speech and Language Technology, Part VI - Language Resources and Language Engineering, Nuria Gala, Reinhard Rapp and Gemma Bel-Enguix (eds.), Vol. 48, New York, USA, pp. 505-543.

Cristea, D., Scutelnicu, A., Pădurariu, C., Boghiu, Ș. (2018). Activity A1.3: Functional and architectural design of the infrastructure that will house the resources and the processing and accessing tools of the consortium; a prototype (in Romanian), internal research report, ReTeRom-CoBiLiRo, RACAI-UPB-UTCN-UAIC.

Cristea, D., Haja, G., Moruz, A., Răschip, M., Patrașcu, M.I. (2011). Partial statistics at the end of the eDTLR project - The Romanian Language Thesaurus in electronic format (in Romanian), in R. Zafiu, C. Ușurelu, H. Bogdan Oprea (eds.) Romanian language. Aspects of linguistic variation. Proceedings of the 10th Colloquium of the Romanian Language Department (Bucharest, 3-4 Dec. 2010), vol. I, Grammar and phonology, lexicon, semantics, terminologies, Romanian history, dialectology and philology, University of Bucharest Printing House, pp. 213-224, ISBN 978-606-16-0046-5.

Georgescu, A., Cucu, H. and Burileanu, C. (2019). Progress on automatic annotation of speech corpora using complementary ASR systems. In *Proceedings of the 42nd International Conference on Telecommunications and Signal Processing (TSP)*, Budapest, Hungary, pp. 571-574.

Li, Ai-jun and Zhi-gang, Yin (2007). Standardization of Speech Corpus. In *Data Science Journal*, vol. 6, supp, 18 November.

Mapelli, V., Arranz, V., Kamocki, P., Mazo, H., and Popescu, V. (2018). New Directions in ELRA Activities. In *Proceedings of LREC 2018*.

Oostdijk, N. (2000). The Spoken Dutch Corpus. Overview and first evaluation. In M. Gravilidou, G. Carayannis, S. Markantonatou, S. Piperidis & G. Stainhaouer (Eds.) *Proceedings of the Second International Conference on Language Resources and Evaluation- LREC-2000*, vol. II, pp. 887-894.

Patrașcu, M.-I., Clim, M.-R., Haja, G. and Tamba, E. (2016). Romanian Dictionaries. Projects of Digitization and Linked Data. In Diana Trandabăț, Daniela Gîfu (Eds.) *Linguistic Linked Open Data*. 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop, Sibiu, Romania, July 13–25, 2015. Revised Selected Papers, Springer, pp. 110-123.

Polat, H. and Oyucu, S. (2020). Building a Speech and Text Corpus of Turkish: Large Corpus Collection with Initial Speech Recognition Results. *Symmetry*, Volume 12 (2), 290.

Sperberg-McQueen, C.M. and Burnard, L. (2018). Original editors, revised and expanded under the supervision of the Technical Council of the TEI Consortium. TEI P5: Guidelines for Electronic Text Encoding and Interchange, version 3.3.0, last update: 31st January 2018, revision: f4d8439.

Stan, A. and Giurgiu, M. (2018). A Comparison Between Traditional Machine Learning Approaches And Deep Neural Networks For Text Processing In Romanian. In *Proceedings of the 13th International Conference on Linguistic Resources and Tools for Processing Romanian Language*, 22-23 Nov., Iași, Romania.

Trandabăț, D. (2018). Activity A1.2. Inventory of available Romanian language data collections at partners or third-party coalitions. Internal research report, ReTeRom-CoBiLiRo, RACAI-UPB-UTCN-UAIC.