THE COBILIRO PROJECT: BUILDING AND DISTRIBUTING A BIMODAL CORPUS FOR ROMANIAN LANGUAGE

DAN CRISTEA[1,2], CRISTIAN PĂDURARIU[1,2], ŞERBAN BOGHIU[1], DANIELA GÎFU[1,2], MIHAELA ONOFREI[1,2], DIANA TRANDABĂŢ[1], IONUŢ PISTOL[1], ANCA BIBIRI[3]

[1] *"Alexandru Ioan Cuza" University of Iasi, Faculty of Computer Science*
[2] *Romanian Academy, Institute of Computer Science - Iaşi Branch*
[3] *"Alexandru Ioan Cuza" University of Iasi, Department of Interdisciplinary Research in Social Sciences and Humanities*

Abstract

CoBiLiRo (*Corpus Bimodal pentru Limba Română* - Bimodal Corpus for Romanian Language) is an on-going research project aimed to collect, standardise and make available a collection of Romanian language files containing both text and audio recordings, aligned at boundaries of sentences, words, phones, and/or other linguistic levels. This paper describes the current efforts carried out as part of this project. We present the design of the format aimed to serve as an annotation standard for bimodal resources, the main operations of the web portal which hosts the corpus, and the automatic conversion flow that brings the inputted file at the format accepted by the Portal.

Key words — bimodal corpus, annotation standard, web portal, speech and text processing, metadata of linguistic resources.

1. *Introduction*

CoBiLiRo is a component part of ReTeRom, a project aiming to push forward the state of the art in Romanian language technology, grouping researchers from four natural language processing laboratories[1] 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. CoBiLiRo aims to create a thesaurus with audio and textual resources annotated on different levels of acoustic and linguistic achievement, which shall stand as the most important reference of this type for the Romanian language, addressing future developments of human-machine interfacing technologies. As such, the project makes a careful inventory of existing bimodal resources at partners, finds ways to harmonize the representation, the annotation and the metadata formats, designs and implements an infrastructure that will finally house the resources, and does a wide dissemination of the bimodal corpus, for research valorisation and use in applications.

[1] RACAI - the "Mihai Drăgănescu" Research Center on Artificial Intelligence of the Romanian Academy in Bucharest, the Speech Processing Laboratory at the Faculty of Electronics, Telecommunications and Information Technology of Politehnica University of Bucharest, the Speech Processing Lab at the Department de Communications, Faculty of Electronics, Telecommunications and Information Technology of the Technical University of Cluj-Napoca, and the Natural Language Processing Group at the Faculty of Computer Science of the "Alexandru Ioan Cuza" University of Iaşi.

In the meaning of the ReTeRom project, but also sharing other views (Mihăilă and Mekhaldi, 2009), by a bimodal corpus we understand a collection of oral records accompanied by their transcripts and their corresponding metadata. A bimodal corpus is hosted on a specialized platform, together with its web access, maintenance services, processing technology and possible applications.
Stages in the evolution of linguistic corpora included: the first generation, which contained only texts (e.g. British National Corpus[2]), the second generation, which contained both oral and written texts (e.g. Michigan Corpus of Academic Spoken English[3]), and the third generation, which included also the alignment between written and oral components. Recent stages in the evolution of corpora are characterized by the inclusion of the video dimension (Rasso and Mello, 2014: 29). The object of our interest is the third generation of corpora.

In order to advance Romanian technologies integrated with natural language processing, our goal is to construct bimodal corpora which prove to be useful in text or recording modelling. Part of our bimodal corpora is annotated on different linguistic levels of the bimodal corpus generated within COBILIRO based on some conventions, which will then be taken as a model for an automatic process, the proposal of an inventory of unstructured data and specifications for the realization of user interfaces.

The next sections will briefly describe existing approaches of bimodal resources, the CoBiLiRo annotation format, the design and implementation of converters, and the CoBiLiRO web platform.

2. Bimodal resources

Research on automatic speech recognition have grown dramatically since the 1960s (Halle and Stevens 1962; Denes & Mathews 1960; Denes, 1960), although the use of oral corpora as a language storage device that should be interpreted or generated by the machine is much more recent. In order to make available these corpora, specialized interfaces have begun to be created. A bimodal corpus should be hosted on a specialized platform, together with its web access, development and maintenance services and applications, where the researcher can find methods and algorithms for corpus use and, from where, in some cases, examples of applications of the corpus in training and evaluation of the technology can be downloaded.

Even if the terms speech corpus and oral corpus are sometimes differentiated (Llisterri, 1996), in English they are often used interchangeable, without a clear distinction between them. In the next paragraphs of this section we will show how these two categories of corpora can be separated.

Speech corpus, in general, means a database of audio files and their textual transcripts, in a format that can be used to create acoustic models that become the engine of speech recognition research. An example is the Switchboard transcripts reviewed at the Institute for Signal and Information Processing (Godfrey and Hollman, 1997). Audio files and their transcripts can be aligned at phoneme, syllable and word level respectively. Within speech recognition systems, prosody models are mainly used to predict prose events (syntactic and semantic accents) associated with a text. It is of interest for research in which the emphasis is on the audible signal, on its

[2] http://www.natcorp.ox.ac.uk/
[3] https://www.lib.umich.edu/database/link/11887

acoustic properties, and the articulatory properties of the vocal tract. The symbolic representation in this case is made using the phonetic alphabet.

On the other hand, the oral corpus is of particular interest for the researches dedicated to the use of a language and the characteristics of the various linguistic levels: lexical, morphological, syntactic, semantic, discourse analysis, conversation, pragmatic studies of communication, sociolinguistics, dialectology, etc., but also in speech technology, for both comprehension and speech generation systems. The transcript marks, in an enriched spelling system, various auditory phenomena that accompany the pronunciation: murmurs, pauses, coughs, laughter, etc. sometimes, doubled by a phonetic alphabet, and the speech recording is aligned with its textual interpretation. Ideally, both language researchers and natural language technology researchers should use the same data set, regarding data collection, transcription, coding and annotation.

3. Design criteria in building the hosting Portal and the adoption of the corpus format

To fulfill CoBiLiRo's purposes, we designed a portal capable to host linguistic resources of the Romanian language, intended to be used for the development of automatic speech recognition and synthesis systems. Many of these resources, used to train acoustic models, are speech corpora containing recordings of different speakers, paired with their corresponding textual transcripts.

In order to ensure naturalness in automatic speech systems, the recordings to be placed at the base of learning experiments should generally be acquired in spontaneous interactions between speakers, therefore readings in lab conditions and eBooks containing actors' voices are less recommended. On the other hand, production of linguistic resources of this type is expensive, both in terms of time and cost. For these reasons, the interest of the ReTeRom consortium was focused towards rich sources of real life speech, audio corpora available online and in the media: radio and television shows, recordings of public meetings for certain institutions, and ad-hoc interviews addressed to people on the street intended to evidentiate local accents and dialectal expressions. As not all of these resources have intrinsic textual transcriptions, we had to prepare the ground for transcribing parts of the corpus, which is by no means a trivial task. Adding transcriptions can be done manually, by listening to the recordings and simultaneously writing down the related text, or by using already existing automatic speech recognition systems. The apparent vicious circle (of using automatic systems to transcribe naturally produced speech, followed by training recognition systems out of the parallel corpus produced this way) is broken down by involving the use of several architecturally different speech recognition systems, which are supposed to make non-symmetrical errors, keeping as correct identical transcribed spans and manually correcting the regions which display discrepancies.

As such, recently, the CoBiLiRo Portal frontend has been upgraded to accommodate the process of uploading bimodal corpora files (speech plus text) even at different moments of time, as each textual transcript can be decoupled from its speech component. Properly annotated, these two components can be paired later, on the Portal, when both are uploaded there. The alignment is assured through segmentation clues, which can be placed at sentence, word or even letter/phoneme boundaries, as will be explained in the following section. Moreover, the frontend

includes functionalities that allow online editing of the two components in view of creating the speech-text alignment, more precisely the inclusion and synchronisation of boundary markers.

The CoBiliRo annotation format is inspired by the TEI-P5[4] standard (Sperberg-McQueen and Burnard, 2018) , the well known scheme for representing a diversity of document types, but also includes elements from other proposals to best fulfil our goals (Romary *et al*., 2017), (Li and Yin, 2007). This standard has been simplified in some aspects and augmented in others to best accommodate the requirements of our bimodal corpus of speech and text data.

The CoBiLiRo format includes a header, which encapsulates metadata related to the resource. This section holds information about: the source of the object stored, the identity of speakers (in conditions of respecting confidentiality terms), the type of voice (spontane or voice-in-reading), aspects regarding technical conditions of the recording, its duration, the type of file stored (*mp3* or *wav*), the segmentation level, etc. The most common level of segmentation is the sentence, but voice can also be segmented in morphological units (words), phonological (phonemes), prosodical (pitch, raise and decrease of the fundamental frequency), or syntactic (nominal group, clause, etc.). These pieces of information are stored in appropriate *xml* tags and attributes, within the *teiHeader* tag.

The segmentation and alignment of the resource is available in *unit* tags, which can be of three types, depending on the manner the resource is represented. The first type ("*file*") is for resources which are held in multiple files. So, for this case, each *unit* tag will have a child node called *speech*, which indicates the name of the file containing the speech component and a child node called *text* containing the textual transcription associated with the audio file specified.

The second type of segmentation is called "*start-stop*" and is used for those resources that present just one speech file, segmented and aligned at temporal boundaries, the text being reproduced between each two such consecutive markers. So, the *unit* tag will contain a *speech* subelement with two attributes *start* and *stop* (in seconds). Along with the *speech* tag, the *text* tag contains a reproduction of the text being spoken. Other components can be added in each *unit* group under specific tags (see an example in the next section).

The third type of segmentation is *"file-start-stop"*, which is a combination of the two types presented above. It is meant to accommodate those resources that contain multiple audio files and a *"start-stop"* segmentation for each of them. So, for each *unit* tag pointing to a speech file, a series of child nodes called *subunit*s are also created. Each *subunit* will hold the *"start-stop"* segmentation, similar to the one described above.

4. Designing and implementing converters

In the bunch of resources contributed by different partners of the ReTeRom project we have identified three specific types of formats, according to which we have designed the first set of convertors, supposed to "understand" the corresponding files and on which they act accordingly to transform them to the CoBiLiRo standard.

[4] http://www.tei-c.org/

The first format is composed of groups of four files: a *wav* file - containing the audio recording, a *txt* file - containing the text associated with the recording speech, a *lab* file - containing the same text as the *txt*, but from which the punctuation has been eliminated and all letters are reduced to lowercase, and a *phs* file - containing a list of all letters present in the recording along with their start-stop moments. The conversion of this format to the CoBiLiRo standard starts with the creation of the header containing the metadata. Part of the information that fills in the header should be inputted by the contributor by following the imposed form provided by the interface. This type of resource is converted to the *file-start-stop* standard representation presented above. All files belonging to the same group will have the same name. After grouping the files, four subunits are created: *speech*, *text*, *lab* and *phs*. Their contents are extracted from, respectively: a *wav* file - containing the recorded segment of voice; a *txt* file - containing the textual transcription on the segment; a *lab* file - containing the same text but in only lower case letters and without punctuation; a *phs* file - containing the sequence of letters in the segment, each paired with time marks showing its start and end as it is pronounced in the *wav* file.

The second format is called MULTITEXT/TEI and is composed of some audio files and an *xml* file containing metadata (not relevant to the scope of our platform) and a series of *div* tags mapping text to the audio files. The first step of the conversion, as in the previous case, is the creation of the CoBiLiRo header and it is done in the same manner as for the first format. Considering that there are multiple audio files in this type of resources, the *"file"* representation is used. The next step is to identify the *div* tags that contain the mapping of the text to the audio files from the original *xml* file. Then the texts in-between consecutive *div* tags are extracted and inserted into *text* tags belonging to different *units.* A *div* tag also contains an *url* tag, where the name of the audio file associated with the corresponding text can be found. This information is inserted into the *speech* tag of the output format belonging to the appropriate *unit* element. As such, the expected pairs of *xml* elements <*speech/*> - <*text/*> are formed.

The third format discussed here is called TEXTGRID and it contains groups of three files. The first type of file is an audio recording that contains the speech part of the resource. The *textgrid* file contains tuples of values *(letter, xmin, xmax)* referring to letters extracted from the text and the time interval between which each letter was spoken. The third file (*txt*) contains information about the energy of the enunciation of each letter, expressed in decibels and the speech frequency. After copying the header information, *unit* tags for each of the audio files are created, with the attribute *speechFile* containing the name of the audio file. Next, a series of child nodes, each containing a sub-element called *speech* and receiving the attributes *start* and *stop* are created. These attributes' values will represent the *xmin* and *xmax* values from the tuples present in the *textgrid* file. The *letter* values from the tuples will be placed in each *subunit* under the tag *text.*

5. The CoBiLiRO web platform

In order to provide a unified platform where all users can upload, store and find resources, we have created a web platform which facilitates collaboration. The platform is available for all CoBiLiRo users that have an account and a password. It integrates the roles of *Admin*, *Contributor* and *Trustee*.

The *Admin* controls the list of users and their credentials and can get information, through logs, on the flow of data on the Portal. This person also manages the creation of accounts from requests addressed by unregistered users.

A *Contributor* may upload its own resources. A user can gain this quality when she/he makes the first request to upload a resource to the Portal[5]. After a resource has been uploaded, the platform processes the content and, provided its format is compatible with one it knows already (as explained in the previous section), it creates one or more *xml* files that concentrate the content of the source files and one header file that includes all metadata that could be automatically extracted. Other information that complements the metadata need to be manually inputted.

Finally, a *Trustee* is a user that has the right to access resources, downloading them or only browsing their metadata.

All the roles described above can access statistics about the types, formats and number of records stored on the Portal.

In order to create an intuitive and friendly User Interface we used Razor, an ASP.NET programming syntax used to create dynamic web pages, and jQuery, a JavaScript library designed to simplify HTML DOM tree traversal and manipulation.

Metadata is client-side validated by checking that all required input fields are submitted. Server-side validation is done by assigning a format type and checking if specific conditions are satisfied for each type of input file. For data persistence, we have used a relational database named MariaDB. As an ORM[6], in order to query our data easier, we used the Pomelo Entity Framework. Our project repositories are hosted on GitHub using Git for source control. In order to provide a high performance and cross-platform application we used the ASP.NET Core framework written in the C# programming language. Another reason we used this technology was that it provides us with data security (protection against SQL Injection and Cross-site request forgery) and a way to secure REST APIs using JSON Web Token. The website is hosted on premises on a CentOS Linux machine.

Our main concern in design was related to efficiency, since we tried to reduce at maximum the response time of each upload request, including format conversion runs.

6. Conclusions and future work

The Portal described hosts at present more resources contributed by the ReTeRom project partners from Bucharest, Cluj and Iași. Their different research interests (speech analysis, speech synthesis, recordings containing prosodic clues and linguistic enquiries from different Romanian provinces) made up the initial bunch of formats for which we have built converters. The designed standard tried to find an equilibrium between these diverse needs.

At the moment, the Portal is ready to host resources observing the types of formats described, but we are open to find other formats and write converters accommodating them as well. What is still to come is the other facet of the transformation technology, convertors receiving in input the CoBiLiRo standard and exporting any of the ones we considered in input. When this will be done, CoBiLiRo

[5] For security reasons this status can presently be trusted only to members of the ReTeRom consortium.
[6] a programming technique for converting data between incompatible type systems using object-oriented programming languages

will become a real hosting-island platform, as an intermediate hub for text2speech and speech2text research technologies.

Through the Portal, a user can also trigger text processing operations, by calling the web NLP technologies installed on the RELATE Platform (Păiş *et al*., in this volume). Other advanced speech processing operations will be integrated in the processing chain as soon as they will reach technological maturity in the ReTeRom project. This way the CoBiLiRo-RELATE tandem will become a research hub for advanced speech and text processing addressing the Romanian language.

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.

References

Mihăilă, C. and Mekhaldi, D. (2009). Bimodal Corpora Terminology Extraction: Another Brick in the Wall. In *International Conference RANLP 2009 - Borovets*, Bulgaria, pp. 236-240 (https://www.aclweb.org/anthology/R09-1044/).

Rasso, T. and H. Mello (eds.) (2014). *Spoken Corpora and Linguistic Studies*, John Benjamins.

Halle, M., & Stevens, K. (1962). Speech recognition: A model and a program for research. In *IRE Transactions on Information Theory*, 8(2), pp. 155-159.

Denes, P. & Mathews, M (1960). Spoken Digit Recognition Using Time-Frequency Pattern Matching. In *The Journal of the Acoustical Society of America*, 30, pp. 1450-1455.

Denes, P. (1959). The Design and Operation of the Mechanical Speech Recognizer at University College, London. In *Journal of the British Institute of Radio Engineer*, 19, pp. 219-229.

Llisterri, J. (Ed.). (1996). Preliminary recommendations on spoken texts. In *EAGLES Document EAG-TCWG-CTYP/P*, May 1996.

Godfrey, John J. & Edward Hollman (1997). Switchboard I, Release 2. In *Linguistic Data Consortium*, Philadelphia.

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

Khemakhem, Mohamed and Foppiano, Luca and Romary, Laurent (2017), Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In *Electronic Lexicography*, eLex 2017.

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