CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis

Carmen García-Mateo*, Antonio Cardenal*, Xosé Luis Regueira, Elisa Fernández Rei**, Marta Martinez*, Roberto Seara*, Rocío Varela*, Noemí Basanta****

* AtlantTIC Research Center
Escola de Enxeñaría de Telecomunicación
Universidade de Vigo
36310 Vigo (Spain)

** Instituto da Lingua Galega
Praza da Universidade, 4
University of Santiago de Compostela
15782 Santiago de Compostela (Spain)

E-mail: carmen.garcia@uvigo.es, cardenal@gts.uvigo.es, xoseluis.regueira@usc.es, elisa.fernandez@usc.es

Abstract

This paper describes the CORILGA ("Corpus Oral Informatizado da Lingua Galega"). CORILGA is a large high-quality corpus of spoken Galician from the 1960s up to present-day, including both formal and informal spoken language from both standard and non-standard varieties, and across different generations and social levels. The corpus will be available to the research community upon completion. Galician is one of the EU languages that needs further research before highly effective language technology solutions can be implemented. A software repository for speech resources in Galician is also described. The repository includes a structured database, a graphical interface and processing tools. The use of a database enables to perform search in a simple and fast way based in a number of different criteria. The web-based user interface facilitates users the access to the different materials. Last but not least a set of transcription-based modules for automatic speech recognition has been developed, thus facilitating the orthographic labelling of the recordings.

Keywords: Galician speech repository, annotation tool, CORILGA

1. Introduction

Galician is a Romance language spoken in the north-western part of the Iberian Peninsula and one of the co-official languages in the Spanish region of Galicia. Galician is one of the so-called minority languages, recognised as such by the Council of Europe in the European Chapter for Regional or Minority Languages. The importance of these languages is attested by the fact that they are spoken in total by more than forty million citizens in the EU. One of the major conclusions of the MetaNet whitepaper (García-Mateo, Carmen and Arza, 2012) is that Galician is one of the EU languages that needs further research before highly effective language technology solutions can be implemented. This development of high-quality language technology for Galician is critical and of an utmost importance for the preservation of a minorized and a minority language such as Galician.

For this purpose, oral resources must first become available. A growing need to compile linguistic resources for Galician has been acknowledged, since although there exists a large amount of speech recordings, these have important flaws and gaps, among others:

- a limited number of recordings from urban settings and of young people
- very few conversations
- few good quality recordings for phonetic analysis
- scarcity of audio books

Furthermore, the materials are only partially transcribed, and the transcriptions are quite fragmented and not publicly available, thus not allowing global searches.

The CORILGA ("*Corpus Oral Informatizado da Lingua Galega*") project tries to fill these gaps with three main aims:

- to collect all the existing material and to compile new speech recordings
- to annotate the speech material at different linguistic levels
- to integrate the data in a single repository that allows structured search.

In fact, the search interface is one of the major outcomes of the project. This project will allow us to conduct further research on different fields, including Galician phonetics, syntax, pragmatics, etc., in addition to their interrelationships.

In this context, the main aim of the paper is to describe the CORILGA corpus for Galician, including some of the tools which have already been developed to process its linguistic resources. The next section is concerned with the main features of the material available in the corpus and with the software tool to access the corpus.

2. CORILGA

The goal of CORILGA project is to release a large high-quality corpus of spoken Galician from the 1960s up to present-day, including both formal and informal spoken language from both standard and non-standard varieties, and across different generations and social levels. The corpus will be available to the research community upon completion.

This corpus, by its nature, will allow conduct research works about different aspects of phonetics, prosody and grammar, and especially the interface between different levels (prosody and syntax, for example). Moreover, CORILGA will can be of help for studying sociolinguistic variation (formal and informal speech, rural and urban varieties, etc.), and the language change from the earliest recordings to the most recent ones and from the most conservative to the innovative varieties.

In this sense, our corpus will profit from materials and results of PRESEGAL, the Galician corpus of the Panhispanic PRESEEA project, being carried out at the University of Santiago de Compostela, and from the Corp-ORAL, a spontaneous speech corpus for European Portuguese developed at the ILTEC in Lisbon (Freitas & Santos 2010).

2.1 Contents

The corpus currently consists of 98 hours of audio recordings with their corresponding transcriptions. Some of these recordings are taken from pre-existing corpora, namely the corpus of Gustav Henningsen (Vázquez Núñez 2012) (recorded in the 1960s in rural areas, amounting to 104 hours) and the "Arquivo do Galego Oral" (AGO) (more than 2,000 hours, recorded mainly in rural areas), among other corpora. These corpora include talks, broadcast interviews, TV shows (news, series, magazines), literary readings, etc.

Due to the character and aims of these sources, most of our corpus consists of recordings of elicited interviews, while formal speeches represent only 6%, and conversations just over 1%. Regarding age groups, speakers of middle and old age predominate, while young speakers (under 30) do not reach 10% of the total time. Recordings from urban and semi-urban settings represent only 20% of the total amount. Therefore, the planned new recordings are aimed at filling current gaps (young urban people, conversations, etc.), in order to balance the corpus.

Our goal is to get around 600 hours of transcriptions in the medium term, but in the long term the corpus should reach at least 1,200 hours.

At the moment, the whole set of CORILGA's recordings is transcribed orthographically, following criteria developed for conversation analysis, adapted from Payratό (2003). Some of them are also partially annotated in other levels: phonetic transcription, and annotations at prosodic, morphological, syntactic and lexical levels. Annotations for type of speech act and topic are also provided.

We aim at achieving full annotation at all these different levels of linguistic description, in order to enhance the possibilities of linguistic analysis, so that linguistic interfaces between different levels could be studied, such as prosody and syntax, or prosody and turn taking in conversation, for instance.

The annotation is done manually, which is a time consuming task. To ease this job, some tasks are performed semi-automatically, such as morphological annotations, and a tentative phonetic transcription.

2.2 Preprocessing tools

The success of the project very much lies on the ability to add material to the repository. The material should not only be the raw audio file, but also the attached transcription files (orthographic, phonetic, prosodic, and syntactic).

Since the process of manual annotation is rather time-consuming, a set of transcription-based modules for automatic speech recognition has been developed, thus facilitating the orthographic labelling of the recordings. Two types of modules are considered. The first one is designed to generate a time alignment from a manual transcription; the second one is designed to generate an initial transcription without any prior information, or from a partial manual transcription. These modules are hence intended to assist the labelling process, generating a time aligned transcription, both at the phonetic and at word levels.

The inputs in the first module are a text file with a word-level transcription and a file with a phonetic transcription. By using an automatic recognition system, the text file is automatically aligned with the recording, generating an EAF format file (ELAN Annotation Format). The file contains two tiers, one with word-level labels, and another one with phoneme-level labels. The IPA Unicode alphabet (International Phonetic Alphabet) is used for the phonemes. The second TIER is either extracted from the manually annotated phonetic transcriptions (if supplied), or it is generated from the word alignment.

The second module performs the same tasks, although in a totally automated way. This system is intended for cases where an incomplete transcription is available. In this case, the annotated part of the recording is time-aligned, using the system mentioned above. The information extracted at this stage (vocabulary, number and characteristics of the speakers, etc.) is used to improve the automatic transcription of the non-annotated part. Even though the transcription obtained with this procedure is not error-free, the quality is usually good enough to facilitate and speed up significantly the process of manual labelling.

3. Description of the Software Repository

The available material, described in the previous section, is taken from different sources and was collected at

different stages. It is vital to homogenize it by defining a computerized repository that fulfils all the envisaged uses. For that reason, we have designed a database whose structure allows for intelligent search. In addition, we have also created a web-based user interface to facilitate users the access to the different materials. In the following, the main features of both the database and the interface are described.

3.1 The database

A database has been designed in order to handle the aforementioned material. The database is written in MySQL (Ullman, J.D. et al., 2002) language and has the following seven tables: “Speakers” (biographic details), “Recordings” (information about the recording), “Topics” (the subject matter of the recording), “Types” (the genre to which the recording is ascribed), “Users” (information concerning the users enrolled in the system), and finally “Recording_type” and “Recording_topic” (information needed to search through the data). More information about the system can be found in the “User’s Manual”, which will be publicly available once the tool is released.

Speakers table contains information about the persons who talk in the audio files that has a significant influence on the recording characteristics, such as the individual’s study level and age, place where they were born, actual address information, language and/or mother language.

Recordings table stores information about the recording date, the topic that is it about, the environment where it took place and the type of information.

Types and Topics tables contain a list of possible kind of recordings (such as: formal, informal, interview or narrative text, etc.) and a list of possible recording topics (music, roles, holiday days, Christmas, marriage, etc.), respectively.

Finally, Users table stores username and password that allow to signed up users log in the system to add or delete recordings. To increase security the password is encrypted avoiding the access to hackers that may accomplish to break the system and access to the database.

The database will be used with the aim of storing and manage information, enabling search in a simple and fast way based in different criteria such as the recording type, recording year, or speaker age, for example. MySQL allows data interconnection between different tables performing complex searches along all information stored.

3.2 The user interface

The user interface has a client-server architecture. From the computing point of view, the client is programmed in HTML5 (Lubbers et al., 2011), using JavaScript and JQuery library (Resig, 2009) to provide dynamism to the web page and improve user experience, while the server is mainly written in PHP (Ullman L.,2011). This configuration allows anyone to access the database using a web browser.

Regarding functionality, the main goal is to allow users to search across the database with a combination of criteria over the different information layers in the recording, along with the searching criteria regarding the type of speaker and the type of recording. Each recording can have the following information attached:

- Orthographic transcription
- Phonetic transcription
  - Syntax annotations
  - Morphological annotations
  - Prosodic annotations
  - Annotation for type of text

The user interface consists of two main parts. On the one hand, an initial window which displays a form with the filtering criteria to select the files for the search and a button to launch the administration session (Figure 1); on the other, a second window to specify the search patterns and to finally display the results of a particular search (Figure 2).

In the first part of the interface, there is also a link for logging into administrative section of the system. The administrator user is mainly in charge of updating the database information and of uploading new contents to the server. This interface provides a drag and drop system to update these contents easily (Figure 3). Administrator users just need to drag the audio and transcription files from desktop into the browser and they will be automatically uploaded to the server. This feature has been implemented using HTML5 file API. Moreover, administrator users can list and delete the files that are stored in the server.

Once the search results are presented to the user, three options are available: 1) listening to the excerpt of the recording, 2) downloading the ELAN (Brugman &Russel, 2004) file with the excerpt, and 3) downloading the PRAAT (Boersma & Weenink, 2013) file. Moreover, users have the possibility of listening to an extended excerpt for each result to have access to more contextual information.

Downloaded ELAN and audio files are cut in the server side according to the excerpt initial and final time. Then, they are packed into a zip file that will be stored into the client device.

For Praat files, the downloaded data also consists in a zip file that contains the audio excerpt, but it contains a PRAAT script created to run this excerpt and show its waveform. The PRAAT script is also created in the server side.

It is also possible to download multiple ELAN or PRAAT excerpt information at the same time setting the checkboxes shown in the results table and using the multiple download button located in the search results header. In this case, the system will download a single zip file that contains a child zip file for each selected excerpt.

When search results are displayed in the second window, users are allowed to define new filter criteria in order to refine the search. To further refine the search, every excerpt has also a bin icon button which let users to hide those that may not be required in the search results.

When the user defines the search criteria, output tiers that will be displayed in the result table need to be selected. Nevertheless, it is also possible to show or hide the other tiers for one excerpt in the search results using the spread out or contract button.

In order to save the search results, we have additionally implemented an option that lets the user download a spreadsheet with information about the record excerpts and speakers. This information can thus be managed with Ms Excel or any compatible spreadsheet software (Figure 3). The output format for these file is CSV and it can be also opened with a text editor.

The ELAN and PRAAT formats have also been selected in view of their widespread acceptance among the linguistic community.

ELAN format consists in a XML file with specific tags which contain the audio file information and annotations for the different tiers. System searches are performed along all the annotation tags of these XML files looking up patterns specified by users.

Praat scripts are portions of code that try to save time to the users automating sequences of operations. Praat documentation provides a scripting tutorial to create this kind of file. We use it to create multiple scripts that let us save user time of loading the file into Praat interface and show its waveform. Thereby, we give our tool the added value of prepare data to perform quick operations.

4. Conclusions and further work

The main components of the CORILGA corpus and a repository for Galician linguistic resources have been described. The research community will greatly benefit from the completion of this project**,** since a public release of the whole system is expected.

5. Acknowledgements

This work has been supported by the Galician Regional Government (CN2011/019, CN2012/160, CN2012/179), the European Regional Development Fund (ERDF) and the Spanish Government (‘SpeechTech4All Project’ TEC2012-38939-C03- 01).

6. References

Boersma, Paul & Weenink, David (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.56, retrieved 15 September 2013 from http://www.praat.org/

Brugman, H., Russel, A. (2004). *Annotating Multimedia/ Multi-modal resources with ELAN*. Proceedings of LREC 2004, Fourth International Conference on Language Resources and Evaluation. Max Planck Institute for Psycholinguistics, The Language Archive, Nijmegen, The Netherlands. http://tla.mpi.nl/tools/tla-tools/elan/.

Fernández Rei, Francisco (dir.) (2010):*Arquivo do Galego Oral*. Santiago de Compostela: Instituto da Lingua Galega. http://ilg.usc.es/ago/ [11/03/2014]

Freitas, Tiago and Santos, Fabíola (2010): “CORP-ORAL: a spontaneous European Portuguese speech resource”, in Freitas, Tiago: *Estudos de corpora. Da teoria à prática.* Lisboa: ILTEC, pp. 103-109 [paper previously presented at LREC’08]

García-Mateo, Carmen and Arza, Montserrat, (2012) *“O idioma galego na era dixital -The Galician Language in the Digital Age”*, Springer, META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors), ISBN 978-3-642-30798-0. http://www.meta-net.eu/whitepapers.

Lubbers, P., Albers, B., Salim, F. & Pye. (2011). “*Pro HTML5 programming*”, Ed.: Apress, http://www.w3.org/html/

Payrató, Lluís (2003): *Pragmática, discurs i llengua oral. Introducció a l'anàlisi funcional de textos*. Barcelona: UOC

Resig, John (2009). *Secrets of the JavaScript Ninja*. Ed.:Manning Publications http://jquery.com/

Ullman, Jeffrey D., Garcia-Molina, H. and Windom, J. (2002) *MySQL: Database Systems: The Complete Book*. Ed.:Prentice-Hall, http://www.mysql.com/

Ullman, Larry, (2011). *PHP and MySQL for Dynamic Web Sites*. Ed.:Peachpit Press, PHP: http://www.php.net/

Vázquez Rozas, Victoria, et al (2014) *Proyecto para el estudio del español de Galicia.* PRESEGAL http://gramatica.usc.es/proxectos/presegal/

Vázquez Núñez, Sandra (2012). "O Corpus (de textos orais) de Gustav Henningsen e a súa importancia para a investigación lingüística e cultural". X Congreso Internacional da Asociación Internacional de Estudos Galegos (AIEG) 2012. Caerdydd/Cardiff, 12-14 setembro 2012.

Figure1: Initial window of the user interface

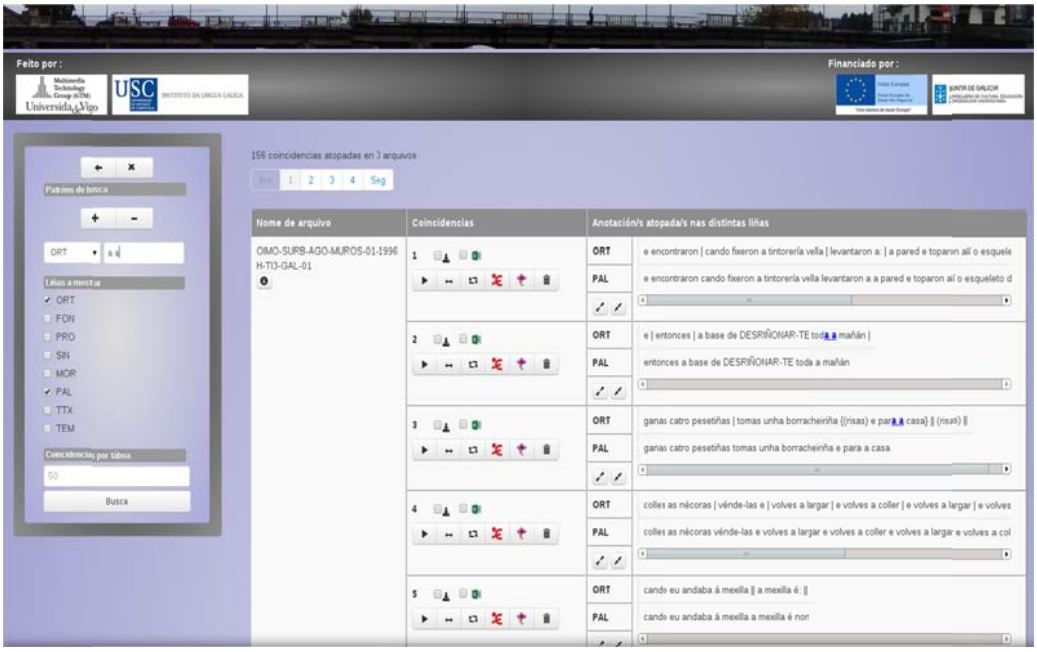

Figure 2: Window with an example of a search output

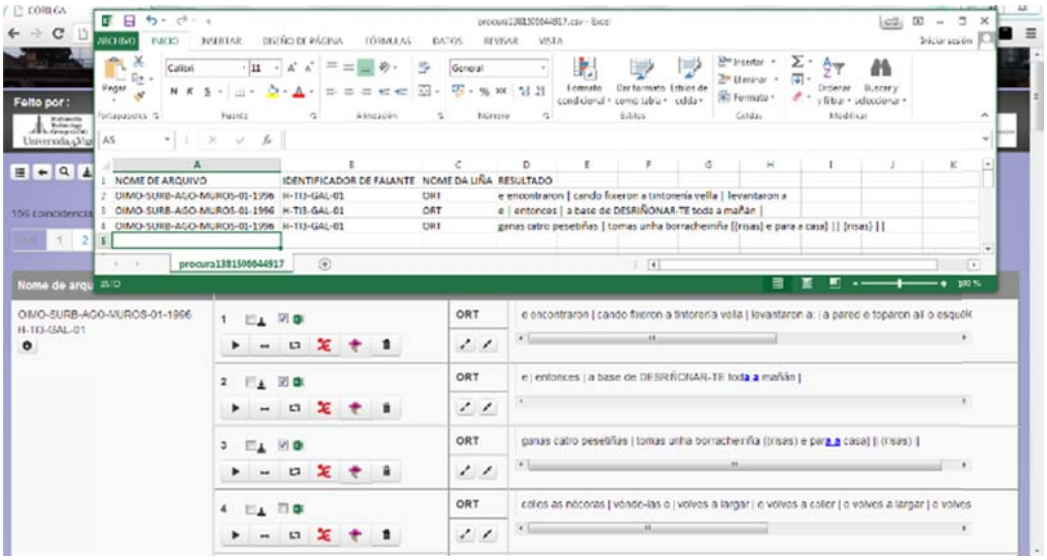

Figure 3: Downloaded spreadsheet with search results information.