
C. García-Mateo / A. Cardenal / X. L. Regueira Fernández / E. Fernández Rei / M. Martínez / R. Seara / R. Varela / N. Basanta Llanes (2014): "CORILGA: a Galician Multilevel Annotated Speech Corpus for Linguistic Analysis". *9th Language Resources and Evaluation Conference (LREC 2014)*. Reykjavik, 26-31 maio 2014.

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**

{carmen,cardenal,mmartinez,rvarela}@gts.uvigo.es
{xoseluis.regueira,elisa.fernandez}@usc.es

*AtlantTIC Research Center, Multimedia Technologies Group, Universidade de Vigo
**Instituto da Lingua Galega, Universidade de Santiago de Compostela

Multimedia Technology Group (GTM)

AtlantTIC

Universidade de Vigo

USC Universidade de Santiago de Compostela

Summary

- Galician is one of the EU languages that needs further research before highly effective language technology solutions can be implemented.
- CORILGA ("Corpus Oral Informatizado da Lingua Galega") 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 across different generations and social levels.
- The corpus will be available to the research community upon completion.
- Its software repository includes a structured database, a graphical interface and a number of 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.
  - A set of transcription-based modules for automatic speech recognition has been developed, thus facilitating the orthographic labelling of the recordings.

CORILGA Corpus

Motivation

A growing need to compile linguistic resources for Galician has been acknowledged, due to the following issues:

- 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
- Materials are only partially transcribed

Goals:

- to collect all the existing material and to compile new speech recordings
- to annotate the speech material at different linguistic levels: orthographically, phonetically, and annotations at prosodic, morphological, syntactic and lexical levels. Annotations for type of speech act and topic.
- to integrate the data in a single repository that allows structured search.

Contents:

- Presently: 98 hours of audio recordings with their corresponding transcriptions
- Material from pre-existing corpora: Talks, Interviews, TV shows, Literary readings
- Speakers of middle and old age predominate, while young speakers (under 30) do not reach 10%
- Recordings from urban and semi-urban settings represent only 20%
- Planned new recordings are aimed at filling current gaps
- Goals: medium term 600 hours, long-term 1.200 hours

Preprocessing Tools:

Two types of transcription-based modules for automatic speech recognition:

- The first one is designed to generate a time alignment from a manual transcription. Inputs: a text file with a word-level transcription and a file with a phonetic transcription.
- The second one is designed to generate an initial transcription without any prior information or from a partial manual transcription.

Software Repository: Database and Web-based User Interface

- Database is written in MySQL with seven tables: "Speakers", (biographic details), "Recording" (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).
- User interface has a client-server architecture
  - The client is programmed in HTML5 using JavaScript and JQuery library.
  - The server programmed in PHP
- 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 may have the following information attached:
  - Orthographic transcription
  - Phonetic transcription
  - Syntax annotations
  - Morphological annotations
  - Prosodic annotations
  - Annotation for type of text

FIGURE 1: Initial window of the user interface which displays a form with the filtering criteria to select the files for the search and a button to launch the administration session

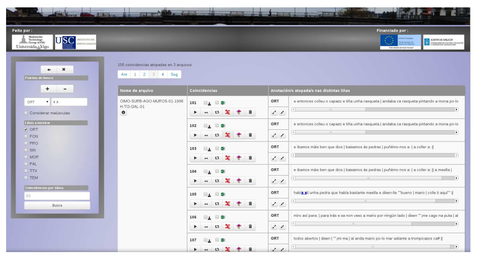

FIGURE 2: Window with an example of a search output

- Once the search is performed, a number of actions are possible:
  - Listening to the excerpt of the recording
  - Downloading the ELAN file with the excerpt
  - Downloading the PRAAT file.
  - Downloading of a spreadsheet with information about the record excerpts and speakers.

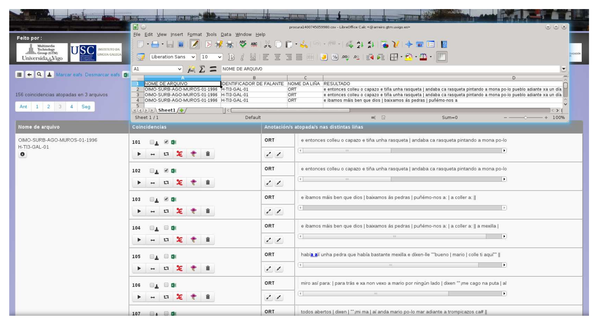

FIGURE 3: Downloaded spreadsheet with search results information

Acknowledgements

This work has been supported by the European Regional Development Fund, the Galician Regional Government (Consolidation of Research Units: CN2011/019, AtlantTIC CN2012/160, TecAnDaLi CN2012/179) and the Spanish Government ('SpeechTech4All Project' TEC2012-38939-C03-01).

SpeTech4All

Unión Europea
Fondo Europeo de Desarrollo Regional
"Una manera de hacer Europa"

Gobierno de España
Ministerio de Economía y Competitividad

Xunta de Galicia