# An Overview of the Existing Language Resources for "Gallego"

Carmen García-Mateo(*) and Manuel González-González(**)

(*) E.T.S.I. de Telecomunicación - Universidade de Vigo, Spain, carmen@tsc.uvigo.es
(**) Instituto da Lingua Galega - Universidade de Santiago, Spain, fg251051@usc.es

## Abstract

*Galician is one of the two official languages in Galicia. Approximately 7% of the population in Spain can speak Galician and would like to use this language for any kind of communication including of course those involving computers. That means the need of developing speech technology products in Galician.*

*In this paper, we present the research projects conducted in the last years aiming at the collection of speech and text corpora in Galician. Most of them have been designed for linguistic study purposes, but also constitute valuable resources for the developing of speech technology products. Finally, as developers of dialogue systems products we present our plans for future data collection that will allow us to deploy in the next two years real-world speech technology applications.*

## 1. Introduction

"Gallego" or Galician is one of the four official languages in Spain. It is mainly spoken in Galicia (region in the northwest part of Spain), as well in the Asturias/Galicia and Castilla-León/Galicia border regions. Galician is a Romance language with some common aspects with Portuguese and Castillian regarding morphology, lexicon and grammar.

In order to have a clear idea of the area of influence of Galician, let us mention that Galicia has an area of approximately 30,000 $Km^2$ and a population of circa three millions (approximately 7% of the population of Spain). According to the 1991 census, 94% of its population understand Galician and 88% can speak it. The linguistic picture in Galicia is that of a bi-linguistic region where both Galician and Castillian coexist, with a clear domination of Castillian over Galician in many areas of application, i.e: newspapers, radio stations, TV stations, education, justice administration, etc. The regional government, "Xunta de Galicia"[1], among other official institutions, promotes its dissemination and use to every extent of the ordinary life. There is an official TV and Radio station [2] and a newspaper fully written in Galician. Some research centers as "Instituto da Lingua Galega"[3] from the University of Santiago de Compostela and the "Centro Ramón Piñeiro de Investigación en Humanidades"[4] from the "Xunta de Galicia", are devoted to the study and promotion of the Galician language.

The deploying of commercial speech products in Galician is uncertain, mainly due to the low number of potential users. There is however much interest among politicians and researchers in the developing of the speech technology for the Galician language. It is clear to us that without the support from the regional authorities is impossible the developing of all the required resources. A Text-to-Speech Conversion system has been developed with the support of "Centro Ramón Piñeiro de Investigación en Humanidades" and it is ready for its integration in any application. In the last years,

several research projects have been launched to acquire the required data to develop speech technology products. Still much work must be done, before we can say that applications in Galician are available to the final user.

In this contribution, we will list, up to our knowledge, the existing language resources for Galician and, as actual developers of such technology, we will present our plans for the next future in order to be able to deploy real-world working man-to-machine dialogue systems in Galician.

## 2. Text Corpora

Two projects are presently under developing aiming at the collection of text corpora: the "Corpus de referencia do galego actual (CORGA)" that is, Present Galician Reference Corpus [5] from the "Centro Ramón Piñeiro de Investigación en Humanidades", and the "Corpus do galego moderno (CGM)", Modern Galician Corpus, from the "Instituto da Lingua Galega".

Next we describe the main characteristics of these two corpora.

### 2.1 "Corpus de referencia do galego actual (CORGA)"

The CORGA corpus has been collect under the direction of Guillermo Rojo and consists of a collection of modern texts (written or generated after 1975) available to the research community without any cost. Several kinds of texts are considered: essays, oral texts, narrative texts, etc. The actual distribution is as follows: 35 texts from newspapers, 58 texts from theatre plays, 58 narrative texts, 41 from essays and 20 oral texts which makes a corpus of about 12 million words.

The data are formatted in a unified style but keeping the original italic and capital letter marks, since these marks can help to make the right interpretation of the marked word/s.

Along with this project, the developing of a morphological analyzer is also under progress in the "Centro Ramón Piñeiro". This tool will allow an automatic analysis and organization of the corpus.

### 2.2 "O Corpus do Galego Moderno (CGM)"

This project began in 1990 under the direction of Antón Santamarina. The Modern Galego Corpus consists of circa 10 million words excerpted from texts originated since the last third of the XVII century (the first text is dated in 1671: "Entremés famoso sobre da pesca do río Miño" de Feixó Araúxo).

Its objective is two folded: first, to be a selection of the written material in Galician since the XVII century up to now, second, to include texts from the oral literature such that: stories, popular songs, idioms, etc.

The texts are morphologically analyzed using a manual procedure but with the concourse of an computerized lexicon: the "Vocabularioo ortográfico da lingua galega".

Presently, 665 texts have been included in the corpus, where 275 of them have already been analyzed and marked up.

In origin the aim of this corpus was to serve as a tool for lexicographic study, but its possibilities are much more. For instance it can be used in the fields of morphology, grammar, dialectology, standardization, updating of the lexicon, etc.

## 3. Speech Databases

Two corpora haven been collected so far that contain speech in Galician. One of them, VOGATEL, has been designed for speech recognition purposes, while the other, "O arquivo do galego oral" aims to be the archive of the oral tradition.

### 3.1 VOGATEL

VOGATEL [5] is the speech database in Galician collected by the "Compañía Telefónica de España" in collaboration with the University of Vigo during 1996. Its main purpose is to provide data for the training and assessment of speech recognizers in this language. The main characteristics of VOGATEL are:

- telephone speech recorded over the public telephone network,
- approximately 7,000 different speakers (men, women and children),
- orthographic transcribed with some additional information about speaker identity (sex and age range, dialect), noise condition, etc.

The data is organized in five categories: digits, connected numbers, command words, phonetically balanced sentences and free speech.

### 3.2 "O Arquivo do Galego Oral"

This corpus has been collected in the "Instituto da Lingua Galega" under the direction of Francisco Fernández Rei and constitutes an invaluable resource for the linguistic study of Galician. Its consists of two different corpora: the Popular Galician Archive ("Arquivo do galego popular") and the Cultivate Galician Archive ("Arquivo do galego culto").

The Popular Galician Archive consists of a library of oral texts belonging to three different periods:

1. Material recorded between 1974 and 1982. This data resembles the oral reality of Galicia in the places where the recordings took place. There are about some thirty hours of data, orthographically transcribed.
2. Material recorded between 1980 and 1988. These approximately 500 recordings are orthographically transcribed. It shows the dialectal picture of the modern Galician.
3. Material recorded between 1992 and 1998. It is organized depending on the age range of the speaker: a) speakers over 66 years old, b) speakers between 40 and 55, and c) speakers under 25. This latest bin is specially interesting since these speakers have already taken Galician language courses in the school.

So far the Popular Galician Archive stores recordings from 4,000 speakers coming from 2,000 different locations up to a number of more than 1,000 hours long.

The Cultivate Galician Archive completes the Popular Galician Archive since it includes recordings from the academia arena: lectures, conferences, round tables about

politics or social situation. About some 70 hours have been recorded.

The relevance of such a corpus is clear: in one hand it is a witness of the actual way of speaking Galician which allows phonological, syntactic and semantic studies. In another hand, it can be considered as a valuable corpus for the study of the intonation.

## 4. Future Plans

Aiming at the developing of dialogue systems in Galician with applications in the telecommunication market, we are presently involved in the collection of several databases. Some acquisition procedures are already in progress, others will be initiated shortly.

The scenario we have defined is that of accessing an e-mail service through telephone. Thus, we need to integrate several products all together. The three main components are that demand corpora for their developing are:

- A text-to-speech conversion system (TTS) in Galician:

  A corpus for modeling the prosody has already been designed and its recording and labeling is in progress. The corpus includes any kind of intonation pattern one can find in Galician. The main purpose of this corpus is to provide data for generating the prosody model of our TTS in Galician, but it is also a very useful tool for more general intonation studies.

- Speech recognizers:

  New speech databases must be collected for this application that can be used for training the speech recognizers.

- A dialogue manager:

  One of the advantages of this project comes from the deploying of such a system in the real world, with actual users. This will allow us to record actual dialogues between the users and the system that can help to improve the dialogue control. The labeling of the recordings is a very high time consuming task as well as heavy. One point under discussion is how to mark up the recordings in order to get a balance between the required additional information and time and resources devoted to the labeling procedure.

## References

[1] “Xunta de Galicia” home page: http://www.xunta.es

[2] “Radio e Televisión Galegas” home page: http://www.crtvg.es

[3] “Instituto da Lingua Galega” home page: http://www.usc.es/ilgas

[4] “Centro Ramón Piñeiro de Investigación en Humanidades” home page: http://www.cirp.es

[5] Invited paper in this workshop