Galician in the Digital Age: An Overview of Automatic Speech Recognition

José Manuel Ramírez Sánchez, Laura Docío Fernández, Carmen García Mateo

GTM research group, AtlanteTic Research Center, University of Vigo, Spain
jmramirez@gts.uvigo.es, ldocio@gts.uvigo.es, carmen.garcia@uvigo.es

Abstract

This study was carried out under the initial state of the European Language Equality project to report technology support for Europe’s languages. In this paper, we show an overview of the current state of automatic speech recognition technologies for Galician. In addition, we compare, over a small set of Galician TV shows, the performance of two of the most reported automatic recognition system with support for Galician: the one developed by the University of Vigo and the one offered by Google.

Our research shows impressive growth in the amount of data and resources created for Galician in the last four years. However, the scope of the resources and the range of tools are still limited, especially in the actual context of services and technologies based on artificial intelligence and big data. The current state of support, resources, and tools for Galician makes it one of the European languages in danger of being left behind in the future.

Index Terms: Galician, automatic speech recognition, resources, under-resourced languages

1. Introduction

Galician is part of the Romance language family with approximately 1,926,000 speakers and is one of the co-official languages of Spain. Despite having a rich cultural tradition, active public institutions, and a growing interest from its community of speakers, the digital presence of Galician is scarce (limited to less than 0.1% of websites on the Internet). In addition, Galician suffers from a lack of digital tools and services that put their survival in the digital age in danger. However, it is not possible to face this situation without having a more precise vision of the current state of both the resources available for Galician and the current status of Galician language development in the digital environment.

In 2012 META-NET produced a series of white papers about the state of European languages in the digital age [1] and [2]. One of these white papers was about Galician [3], and its results were moderately optimistic about the state of Language Technologies (LT) support for this language. Ten years later, the status of LT changed: Machine Learning have become state-of-the-art in the area, and new players such as Google have entered the field to lead development and research. In this new scenario, the studies carried out in 2012 required an update. European Language Equality (ELE) project [4] carried out a series of studies that extend the results of META-NET.

The ELE project, funded by the European Commission, provides a road-map and framework to achieve full digital language equality in Europe by 2030. One of the phases of ELE was an intense metadata collection of data, resources, and tools of the 24 European official languages and almost 40 regional and minority languages. The data was collected in 2021-2022 and it was used to calculate a metric of the technology support for each language. This metric is called Digital Language Equality, and it takes into account both technological and contextual factors to evaluate the risk of a language in the digital era.

In this paper, we present in detail some of the relevant results discovered under the ELE project for Galician, comment on the Galician’s DLE score, and discuss the challenge of automatic transcription for Galician using real data and state-of-the-art systems.

2. Language Technology for Galician

The META-NET white papers series made in 2012 about the state of LT support for Galician [3] were moderately optimistic. The study concluded that despite an LT community of researchers and a series of state-of-the-art resources and technologies, the scope of resources and the variety of available technologies was very limited compared to the resources and tools for other languages such as Spanish. It was concluded that the Galician LT industry was very small, and it was proposed as the only possible alternative to reverse this situation to make a significant effort to create more and better LT resources for Galician.

Ten years later, the arrival on the scene of new Machine Learning algorithms and models, together with a large corpus of texts and a considerable increase in computing power, has changed the landscape of LT with a remarkable progress. However, these changes have not brought an equal improvement in the quality of services and support for all languages [5] and [6].

The aim of the metadata collection activities made by the ELE project was to discover the components that contribute to the level of technological support of the European languages. The information collected by ELE was stored in its public catalogue the European Language Grid (ELG) 1. These data allow for a cross-language comparison based on a well-defined set of quantifiers, measures and indicators [5] and [6]. In addition to this data collection, an analysis metric and a scale of four levels of technological support for a language2 were defined. The technical details of the DLE metric designed by ELE can be found in [5] and [6]. The scale used, inspired by the one designed in the META-NET technical document series [1], is as follows:

- Weak or no support: the language is present (as content, input or output language) in <3% of the ELG resources of the same type.
- Fragmentary support: the language is present in >=3% and <10% of the ELG resources of the same type.
- Moderate support: the language is present in >=10% and <30% of the ELG resources of the same type.

1https://live.european-language-grid.eu/
2The overall level of support for a language was calculated based on the average coverage in all dimensions investigated.
Preliminary Results
European Language Equality
Results based on raw counts of the 11,000+ language resources and language technologies currently described with metadata records in the ELG platform.

- Good support: the language is present in $\geq 30\%$ of the ELG resources of the same type.

Figure 1 visualises the main findings of the ELE’s cross-language comparison. The best supported language is English and it is the only language ranked in the good support group. Among the Iberian languages, Galician together with Basque hardly exceed the threshold of the lowest level. Although it must be taken into account that some contextual factors, such as its quality of co-official language, raise its total DLE score. Catalan occupies the best position within the co-official languages, belonging to the group of languages with fragmentary support. Spanish is inside the group of languages with moderate support, but it is similar to English in some dimensions (in terms of available models).

One of the factors that considerably influences the DLE score achieved by Galician is the lack of multimedia databases (audio + video) and the small number of acoustic databases available. The DLE metric gives importance to this type of resources because they are useful in many tasks. However, the vast majority of the resources available for Galician are corpus or linguistic resources such as lexicons. In section 2.1 we offer the details of the main sources, distribution, and existing gaps of the resources available for Galician. Section 2.2 will focus on describing the ASR systems found with support for Galician.

2.1. Data and Resources

One of objectives of this research is to show the current state of speech and language technologies in terms of resources, services, and community for the Galician language. To do this, we have carried out an enormous and exhaustive collection of data that has provided a detailed, empirical and dynamic map of the technological support of Galician. We collected the research data used in 2021-2022 under the framework of the “European Linguistic Equality” project funded by the European Commission. During this time we have consulted repositories of public institutions, universities, research centers, and digital platforms to share code, models, or databases. In addition, we contacted scientific and business personnel related to Galician to obtain information or resources of no public domain.

We noticed, in our analysis, an increase in the resources and corpora created between 2018-2021 (67.69% of those indexed). However, tools and services developed in the same period have not increased to the same degree (37.27% of those indexed). There is a significant imbalance in the distribution of resources and corpora by technologies. Table 1 shows that corpora for text resources are the most prevalent, whereas there are few corpora for other technologies.

Most of the resources come from three types of sources: non-Galician public institutions and research centers (42.17%), Galician public institutions (28.92%), and non-Galician private companies or public institutions (28.92%). It is important to note that most of the resources, services, and tools created by
non-Galician entities tend to belong to multilingual projects or products that include Galician as one of several languages. However, those created by Galician entities tend to focus on Galician and therefore offer better quality. Regarding the accessibility and use of resources for Galician, since most of them have been developed by open-source projects, study centers and universities, they can be downloaded and used under licenses that are mostly compatible with GNU/GPL. However, around 20% of the indexed elements are not available for commercial purposes.

The situation of Galician in terms of data and resources is optimistic for most of the technologies that process and use text. However, regarding multimedia data, there is an enormous gap and speech processing technologies seem to be less mature than technologies based on text processing.

2.2. Automatic Speech Recognition for Galician

The status of the LT for Galician is behind the majority of languages, such as English or Spanish. However, in the case of those technologies that use multimedia data, the performance of the systems for Galician is usually even lower [7] and [8]. The amount and low diversity of the data available in Galician are the main obstacles. This lack of resources makes it hard to build models capable of handling the variability of spontaneous speech or the presence of multiple speakers in the audio. However, it is possible to build ASR systems that work in controlled environments with similar conditions to the training data.

The quality of automatic transcriptions in Galician is worse compared to other languages such as Spanish or English [3]. However, we have not found any quantitative study comparing different automatic transcription systems with support for Galician. Using the data obtained from the ELE Galician’s report, we have found that only two automatic transcription systems for Galician are available on the web: a system developed by the University of Vigo (GTM-UVigo ASR) [9] and the Google Cloud Speech-to-Text that is described below [10].

2.2.1. GTM-UVigo ASR

The GTM-UVIGO ASR [9] is a large vocabulary automatic speech recognition (ASR) system built using the Kaldi toolkit [11]. It uses an hybrid Deep Neural Network - Hidden Markov Model (DNN-HMM) for acoustic modeling, and a rescoring of a trigram based word-lattices, obtained in a first decoding stage, with a language model (LM) based on a recurrent neural network (RNN). GTM-UVigo ASR also has a speech activity detection module and uses as audio representation a 40 Mel-frequency cepstral coefficients plus a 100-dimensional i-Vector per frame.

The acoustic model architecture used is a multi-spliced Time Delay Neural Network (TDNN) feed-forward to model long-term temporal dependencies and short-term voice characteristics [9]. The topology of the network consists of an input layer followed by 5 hidden layers with 1024 neurons with RELU activation function [9]. Asymmetric contexts of 13 frames were used in the past, and 9 frames in the future [9].

<table>
<thead>
<tr>
<th>Only text</th>
<th>Multimodal</th>
<th>Only audio</th>
<th>Only video+audio</th>
</tr>
</thead>
<tbody>
<tr>
<td>91.93%</td>
<td>5.66%</td>
<td>1.61%</td>
<td>0.8%</td>
</tr>
</tbody>
</table>

The LM based on n-gram of order 3 was trained using the SRI Language Modeling Toolkit, a modified Kneser-Ney discounting, and a weight interpolation with lower orders. For training the RNN LM, the Kaldi software was also used. The neural network language model topology consists of a RNN with 5 hidden layers and 800 neurons, where TDNN layers with activation function RELU, and Long Short-Term Memory layers are combined. The training is performed using Stochastic Gradient Descent, and 20 epochs. The LMs have been trained with the same material. Figure 2 shows a block diagram of the recognition process.

2.2.2. Google Cloud Speech-to-Text.

The Cloud Speech-to-Text API (S2T) was launched in April 2017 and is one of the cloud services offered by Google Cloud. Its main function is the automatic transcription of audio files. The official documentation describes the service as follows: Accurately convert speech to text with an API powered by Google’s best-in-class AI research and technologies. S2T offers support for 380 languages, including Galician and the co-official Spanish languages (Basque and Catalan). S2T offers several automatic transcription models adapted to different types of audio. The default set of models are available for all languages. The performance of the S2T results depends on four main factors:

- The language.
- The acoustic quality of the recording.
- The S2T model used to perform the transcription.
- The phonetic and linguistic complexity of the phrases present in the audio.

Given the commercial nature of Google Cloud S2T, the details of models used as automatic recognition systems are not...
public. The public information available on the online documentation is that all the models used are deep neural networks trained over Google’s data sets.

3. ASR systems evaluation.

The second objective of this research is to compare the only two ASR systems for Galician presented in section 2.2. The evaluation consisted in calculating and comparing, based on a quantitative metric, the quality of the automatic transcriptions of multimedia files with voice in Galician generated by S2T and by the ASR of the Multimedia Technologies Group of the University of Vigo (GTM-UVigo ASR). For this, a group of programs from 4 Galician production companies was chosen: TVG, Voz Audiovisual, Setemedia, and Adivina. Table 2 shows the final list of selected programs.

The methodology chosen to evaluate the systems consisted of carrying out the standard process used by the scientific community in the area of Speech Technologies to measure the performance of ASR. This methodology uses the word error rate (WER) as a metric. According to the quality standards for automatic subtitles of “Radio Televisión Española” for bilingual territorial news programs [12] a WER equal to or less than 20% is considered acceptable.

The details of the results obtained for all television programs, using the WER as a quality metric, can be seen in Figure 3. We can summarize the results in the following points:

- The ASR-UVigo system produces better quality transcripts than the Google Cloud S2T system in each case.
- The most frequent errors for both systems are of the substitution type. Although, omission errors are frequent in the Google Cloud S2T system.
- The quality of the transcriptions is inversely related to the amount of spontaneous speech, overlapping speech, and the presence of music or other acoustic events in the dialogues.

The results can be grouped into 3 well-defined groups:
- Telexornais of the CRTVG (acceptable quality, 20%)
- Voz Audiovisual programs (poor quality, 60%)
- Setemedia and Adivina programs (unsuitable, 70%)

4. Conclusions

This study shows that in the case of Galician the scope of LT resources and the range of LT tools are still limited compared with other languages such as English or Spanish. LT resources for Galician are not sufficient in terms of quality or quantity to develop state-of-the-art technologies based on data greedy paradigms such as deep learning models.

The current state of automatic transcription systems for Galician is far from the results in other languages. However, it is possible to reach WERs below 20% if the audio comes from a controlled environment with few speakers. There are very few automatic transcription systems with support for Galician. In our research, we only found two: the system developed by the University of Vigo (GTM-UVigo ASR) and Google Cloud Speech-to-Text (S2T). The performance of the GTM-UVigo ASR system over a small data set of TV shows in the Galician is better than the Google S2T system.

Our report shows the necessity to make a substantial effort to create LT resources for Galician, especially multimedia resources. The need for large amounts of data is now more urgent than ever due to the great potential that artificial intelligence and big data can offer. These technologies are already crucial today, and there is a danger for under-resourced languages like Galician to be left behind in the future.

5. Acknowledgements

To all the people who helped us with information or resources during the data collection for the ELE project. Especially to professors Pablo Gamallo, Xavier G. Guinovart and Marcos García González. The work presented in this article was co-financed by the European Union under grant agreement no. LC-01641480 – 101018166.

To the TILGAL project for providing the experimental framework used in section 2. The project “TILGAL: Herramienta de Transcripción y Subtitulado Automático al Galego” is a project coordinated by the “Cluster do Audiovisual Galego” that has been funded by the “Agrupaciones Empresariales Innovadoras. Convocatoria 2022. Ministerio de Industria, Comercio y Turismo del Gobierno de España” and the “Secretaría Xeral de Política Lingüística da Xunta de Galicia”.

We would also like to thank the Consellería de Cultura, Educación, Formación profesional e Universidades of the Xunta de Galicia for the “Centro singular de investigación de Galicia” accreditation 2019-2022 and for the “Axudas para a consolidação e estructuración de unidades de investigación competitivas do Sistema Universitario de Galicia -ED431B 2021/24”, and the “European Union for the “European Regional Development Fund - ERDF”.

<table>
<thead>
<tr>
<th>ID</th>
<th>Channel</th>
<th>TV show name</th>
<th>duration (s)</th>
<th>words</th>
</tr>
</thead>
<tbody>
<tr>
<td>01_SM</td>
<td>Setemedia</td>
<td>talk_show_spont_speech_sm01</td>
<td>4307</td>
<td>8054</td>
</tr>
<tr>
<td></td>
<td></td>
<td>talk_show_sm02</td>
<td>3414</td>
<td>6203</td>
</tr>
<tr>
<td></td>
<td></td>
<td>talk_show_spont_speech_sm03</td>
<td>5981</td>
<td>7277</td>
</tr>
<tr>
<td>02_Voz</td>
<td>Voz Audiovisual</td>
<td>factual_tv_voz01</td>
<td>2109</td>
<td>4025</td>
</tr>
<tr>
<td></td>
<td></td>
<td>game_show_voz02</td>
<td>2149</td>
<td>3807</td>
</tr>
<tr>
<td></td>
<td></td>
<td>game_show_voz03</td>
<td>4316</td>
<td>11380</td>
</tr>
<tr>
<td>03_ADV</td>
<td>Adivina</td>
<td>factual_tv_adv01</td>
<td>3421</td>
<td>6898</td>
</tr>
<tr>
<td></td>
<td></td>
<td>factual_tv_adv02</td>
<td>2951</td>
<td>3837</td>
</tr>
<tr>
<td></td>
<td></td>
<td>factual_tv_adv03</td>
<td>3091</td>
<td>7980</td>
</tr>
<tr>
<td></td>
<td>Telexornais</td>
<td>broadcast_news_tvg01</td>
<td>2820</td>
<td>9990</td>
</tr>
<tr>
<td></td>
<td></td>
<td>broadcast_news_tvg02</td>
<td>2820</td>
<td>7726</td>
</tr>
<tr>
<td></td>
<td></td>
<td>broadcast_news_tvg03</td>
<td>3313</td>
<td>9042</td>
</tr>
<tr>
<td></td>
<td></td>
<td>broadcast_news_tvg04</td>
<td>3631</td>
<td>9826</td>
</tr>
</tbody>
</table>
6. References


