Enhanced CORILGA: Introducing the Automatic Phonetic Alignment Tool for Continuous Speech

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

* AtlantTIC Research Center - Escola de Enxeñaría de Telecomunicación- Universidade de Vigo
Campus Universitario 36310 Vigo (Spain)
** Instituto da Lingua Galega Universidade de Santiago de Compostela
Praza da Universidade, 4, 15782 Santiago de Compostela (Spain)
E-mail: carmen.garcia@uvigo.es

Abstract

The *Corpus Oral Informatizado da Lingua Galega* (CORILGA) project aims at building a corpus of oral language for Galician, primarily designed to study the linguistic variation and change. This project is currently under development and it is periodically enriched with new contributions. The long-term goal is that all the speech recordings will be enriched with phonetic, syllabic, morphosyntactic, lexical and sentence ELAN-complaint annotations. A way to speed up the process of annotation is to use automatic speech-recognition-based tools tailored to the application. Therefore, CORILGA repository has been enhanced with an automatic alignment tool, available to the administrator of the repository, that aligns speech with an orthographic transcription. In the event that no transcription, or just a partial one, were available, a speech recognizer for Galician is used to generate word and phonetic segmentations. These recognized outputs may contain errors that will have to be manually corrected by the administrator. For assisting this task, the tool also provides an ELAN tier with the confidence measure of each recognized word. In this paper, after the description of the main facts of the CORILGA corpus, the speech alignment and recognition tools are described. Both have been developed using the Kaldi toolkit.

Keywords: minority languages, alignment and recognition tool, confidence measure

1. Introduction

The *Corpus Oral Informatizado da Lingua Galega* (CORILGA) (Garcia-Mateo, 2014) is a corpus of oral language for Galician language that was designed to study the linguistic variation and change, but it can also be useful for other purposes (language teaching, linguistic repertoires, etc.).

The corpus consists of spoken language recordings of different levels and registers (standard language, literary, popular, formal and colloquial language, rural and urban) and of different types: conversations, guided interviews, speeches, lectures, recited poetry, theatre and media.

It collects the language of speakers of different ages and both sexes that were recorded from mid-1960 to the present time. Apart from the recordings specifically made for this corpus and the public oral texts taken from the media or the Internet, CORILGA incorporates texts and, in some cases, transcriptions from other corpora, among which they are the projects of the *Arquivo do Galego Oral (AGO)* (Fernández Rei, 2011), *Prosodia da lingua formal*, *AMPER-Galicia* (Escourido, 2008), or the personal corpora, namely those from Gustav Henningsen (Vázquez, 2012), from Francisco Dubert about Santiago de Compostela, from Xosé Luís Regueira about Vilalba region, Manuel Rico including radio interviews, from Noemi Basanta including conversations, from Eduardo Louredo about Leiro, from Miguel Abraira and Xabier Iglesias.

The corpus is currently under development, and it is periodically enriched with new contributions, either recordings, or aligned and annotated texts. For this reason, not all of the transcriptions are fully reviewed and in some instances, incomplete transcriptions and annotations can be found.

The medium-term goal is that all the speech recordings will be enriched with ELAN´s complaint phonetic, syllabic, morphosyntactic, lexical and sentence tiers. Currently, the orthographic transcription is aligned at the phonic sequence level and at word level, and the phonetic transcription at the segment level. Morphosyntactic annotations are also available.

The manual transcription of the speech recordings is a highly time consuming task that is not even error free. This process can be speeded up with the use of an automatic speech recognizer. Therefore, CORILGA repository has been enhanced with an automatic alignment tool, available to the administrator of the software repository that aligns speech with an orthographic or a phonetic transcription. In the event that no transcription is available, a speech recognizer for Galician provides one. This recognized text may contain errors that will have to be manually corrected by the administrator. In order to do so, the tool also provides a tier with the confidence measure of each recognized word. The main goal of this paper is to introduce this tool. Before that, the main characteristics of the software repository are described.

2. Description of the software repository

The repository includes a structured database, a graphical interface and a number of speech processing tools. The use of a database allows simpler and faster search through different criteria. Regarding the software repository, the database is written in MySQL language. The user interface has a client-server architecture. The client is programmed in HTML5, using JavaScript and JQuery library while the server is mainly written in PHP. This configuration allows access to the database using a web browser. There is an administrator user who is mainly in charge of updating the database information and of uploading new contents to the server. A web-based graphical interface allows users to

conduct a search through the material in a structured way. This interface has been redesigned for a better user navigation experience.

With regard to the 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 may have the following information attached:

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

Figure 1 shows an example of the output for a typical search of an orthographic pattern. Once the search results are presented to the user, two options are available: 1) playing the excerpt with or without expanded context, 2) downloading the ELAN (Brugman & Russel, 2004) or PRAAT (Boersma & Weenink, 2013) file with the excerpt.

3. Speech-Recognition-Based Tools

The Kaldi toolkit (Povey, 2011) is used in the alignment and recognition processes. Kaldi is a powerful speech recognition toolkit written in C++, freely available under the Apache License. It supports linear transforms, MMI, boosted MMI and MCE discriminative training, feature-space discriminative training, and deep neural networks. Kaldi aims to provide a software that is flexible, extensible and easy to modify by the users. Afterwards, the training material used for the development of the speech recognizer in Galician is described, as a previous step before the introduction of the alignment and recognition tools.

3.1 Speech Training Material

Multilingual acoustic models have been trained with TCSTAR database (Docio-Fernandez, 2006) which contains European and Spanish Parliament recordings with their respective transcriptions in Spanish, and with Transcrigal database (Garcia-Mateo, 2004) which contains 31 hours of news recordings from regional television with transcriptions in Galician.

Statistical language models play a crucial role in speech recognition systems. The best results are achieved when different language models are combined (Broman, 2005), usually in a linear way. We use a number of language models that have been trained before-hand using different sources of material: news from newspapers and TV stations, material from the Galician Wikipedia, and several manual transcriptions of the material of CORILGA. This way, the language models used in the combining process are: two generic models for Galician with vocabularies of 60 thousand words, a language model trained with the partial transcription of the speech file, and a language model trained with texts of similar files of the CORILGA corpus. The final language model has a vocabulary of around 100 thousand words.

3.1 Alignment and Recognition Tool

The easiest way to get a time-aligned annotation is the case when an ELAN file with an orthographic tier is available. In this case, the tool extracts the initial and final time marks of the sentences from the ELAN file, and then it uses the speech recognizer for aligning each word and phoneme of those sentences.

However this is not often the case. A more complex situation occurs when the transcription file does not include any time mark. Here the process of alignment is divided into two steps. The first one consists in audio segmentation that looks for chunks of audio segments separated by silences. Afterwards, speech alignment is performed for each chunk using a speech recognizer.

For this latter situation, there are two possible scenarios. The difference is the presence or absence of a partial initial transcription. Many of the files in the CORILGA corpus are partially transcribed, usually at the beginning of the audio (around 15 minutes). If available, this text is used to improve the language modelling of the speech recognizer. Therefore, it is used to train one adapted language model that then will be mixed with others (described in the previous subsection). Additionally, this partial annotation is used to select similar material in the remaining CORILGA database that will be used to train a third language model. The hypothesis made here was to assume that the first part of the file was a good representation for the entire file. In this context, it was necessary to make perplexity calculations. The perplexity measures the quality of a language model to recognize an objective text. Defining the partial transcription as the target text, the perplexity reached for a mixture model would be similar if the objective would be the rest of the file. As a consequence, the process for obtaining the best mixtures starts giving initial weights for the input models, and using and iterative algorithm to find the best perplexity and the best weights combination.

After the language model is obtained, the task of recognition continues with a first step of decoding the whole audio. The first step is used to segment the audio file by silence areas. After that, in order to improve the performance of acoustic models, the audio data is modified through FMLR transforms and finally a word transcription is provided.

This recognition output is tagged with the level of confidence in order to perform a manual check of the automatic generated transcriptions. The confidence measure is obtained from the recognition scores using a heuristic algorithm. The tolerance level can be adjusted by the administrator of the application. As a default, three levels of confidence are obtained: good, medium and poor. Figure 2 shows an example of the output in an ELAN file where all the available tiers are displayed. Namely, the recognized text (“Rec-ORT” tier), its segmentation into words (“REC-PAL” tier) with attached confidence level (“REC-CONF” tier), morphological labels (“REC-MOR” tier) and phonetic segmentation (“REC-PHON” tier). The morphological labels are automatically obtained.

This output is intended to serve as an initial annotation for

a manual revision before uploading it to the CORILGA repository.

4. Orthographic Alignment Tool: How to Use it

Figure 3 shows the user interface of the alignment tool where the required files can be either, dragged or browsed. These are an audio file in WAV format and an orthographic transcription file, which can be in an ELAN file (.eaf) or in a .txt file encoded in UTF-8 format. In the event of providing an .eaf file, this must contain an orthographic transcription in one or several ELAN lines (tiers) with its names ending in "-ORT". The lines can be divided in segments (sentences), in this case the alignment is faster and more accurate or there may be only one segment including the entire text. A phonetic transcription file (in .txt format) is optional.

The output will be provided in an .eaf file containing the following lines:

- orthographic (ORT)
- words (PAL)
- morphological (MOR)
- phonetic (PHON)

5. Recognition Tool: How to Use it

Figure 4 shows the user interface of the recognition tool where the required files can be either, dragged or browsed as in the previous tool. There are two possible modes of operation depending on whether a partial word transcription is provided or not. The required input files are an audio file in WAV format for both cases, and optionally an ELAN file (.eaf) partially transcribed. It must contain (at least) the orthographic transcription in one or several ELAN lines (tiers) with its names ending in "-ORT". The lines can be divided into segments (sentences) or there may be only one segment including the entire text.

In both cases the output will be provided in an .eaf file containing the following lines:

- orthographic (ORT)
- words (PAL)
- confidence labels (CONF) (this tier may be a help for the manual correction of the transcription). Its annotations are based on the probability of success of the word being recognised and they can be empty or having the value: "REGULAR" (for medium quality) or "MAL" (for poor quality) . If not label is provided, it means that the interval has a high confidence. Thus, when reviewing the text, we can go directly to the areas where there are annotations, skipping the areas with high probability of success.
- morphological (MOR)
- phonetic (PHON) (this tier may be a help for a future phonetic transcription)

6. How to Use CORILGA: Some Insights

This corpus was primarily designed for linguistic research purposes, mainly to study language variation and change, although it can be used for some other goals. The interface was designed to allow searches for targeted groups of speakers, separated by age, sex, education level, or L1 (Galician or Spanish), targeted genres and types of text (formal or informal, conversation, talk, media, reading, among others) and the year (or group of years) when the file was recorded. Thus, that will facilitate not only researches on language variation between speakers (by urban/rural context, sex, age or education level) and between language varieties (by place, formal/informal, oral/written language, genre and type of text), but studies on language change as well, both in apparent (comparing different generations at a given time) and in real time (comparing the same age groups recorded with a distance of decades).

One example could be research on different linguistic processes that have been observed as varying or undergoing a change in the last decades. Among these, the loss of clitic /no/, replaced by the unmarked form /o/, has been noted in some contexts among young people and in formal styles. This could be assessed by comparing the usage of /no/ and /o/ in the concerned contexts by different groups or in different styles. That can be achieved by searching, for instance, the chain "*non o*" in the PAL tier and [n o n o] in the FON tier, obtaining the maintenance of the clitic pronoun /no/ (spelling "non o", morphological analysis [NEG no(n) + CLIT no]), and the sequence [n o ŋ o] for the change accomplished [NEG non + CLIT o].

Other annotations can be combined to achieve useful results for phonetic, morphological or lexical variation and change. However, for certain purposes some limitations are still present. For example, the fact that the search engine takes the temporal limits of the ORT tier as the time frame of reference makes it difficult to study some phonetic processes, since the system gives positive results when the desired segments or chains are present within the same temporal frame of the ORT annotation, irrespective of the two desired segments are aligned or not. As a case in point, if we want to know if the syllable final [s] undergoes rothatization in some contexts, and we launch a search for "s" in the ORT tier and [r] in the FON tier, we obtain all the results which contain at least one "s" in ORT and one [r] in FON, whether they are aligned or not. That can be partially avoided by using more complex search chains of segments, but some improvements may be done in the future to address this issue.

Besides research in linguistics, the CORILGA corpus can serve other purposes too, such as language teaching (allowing to teach different registers), or as a repository of language and discourse resources useful for writers, screenwriters, translators, teachers, and for other corpus-based applications.

7. Conclusions and Further Work

The recently integrated tools for text and speech alignment have been described. The development of this corpus is an example of how automatic speech recognition tools can facilitate the completion of high-quality linguistic resources.

As said before, the work is still on progress, being two lines of action envisaged: one is to keep on adding more material,

and the other to improve the performance and functionality of the search software and speech processing tools.

8. Acknowledgements

This research was funded by the Spanish Government under the project "Cambio lingüístico no galego actual" (FFI2012-33845), the Galician Government through the research contracts: GRC2014/024 and GRC2013/40 (Modalidade: Grupos de Referencia Competitiva 2014), the "Rede Tecnoloxías e Análise dos Datos Lingüísticos (TECANDALI)" and "AtlantTIC Project' CN2012/160, and the European Regional Development Fund (ERDF).

9. Bibliographical 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/

Broman S. and Kurimo M. (2005), "Methods for Combining Language Models in Speech Recognition", in Proc. Interspeech, pp. 1317-1320.

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/.

Docio-Fernandez L., Cardenal-Lopez A., and García-Mateo, C. (2006): TC-STAR 2006 Automatic Speech Recognition Evaluation: The UVIGO System. In: *TC-STAR Workshop on Speech-to-Speech Translation*.

Escourido A., Fernández Rei E., González González M., Regueira Fernández X. L. (2008): "A dimensión prosódica da oralidade: achega dende AMPER". In E. Fernández Rei & X. L. Regueira (eds.): *Perspectivas sobre a oralidade*, Santiago de Compostela: Instituto da Lingua Galega / Consello da Cultura Galega, pp. 75-93.

Fernández Rei, F. (2011): "O arquivo do Galego Oral do Instituto da Lingua Galega", *A Trabe de Ouro*, 86, pp. 295-298.

Garcia-Mateo C., J. Dieguez-Tirado, A. Cardenal-Lopez, and L. Docio-Fernandez. (2004) "Transcrigal: A bilingual system for automatic indexing of broadcast news". In *Proc. Int. Conf. on Language Resources and Evaluation*, volume 6, pages 2061–2064, Lisbon, Portugal, May.

García-Mateo C. & 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". In *Proc. 9th Language Resources and Evaluation Conference (LREC 2014*). Reykjavik, 26-31 May 2014.

Povey D., Ghoshal A., Boulianne G., Burget L., Glembek, O., Goel N., Hannemann M., Motlicek P., Qian Y., Schwarz P., Silovsky J., Stemmer G., Vesely K, (2011), "The Kaldi Speech Recognition Toolkit", In *Proc. of IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. Hawaii, US. IEEE Catalog No.: CFP11SRW-USB

Vázquez Núñez, S. (2012): "O Corpus (de textos orais) de Gustav Henningsen e a súa importancia para a investigación lingüística e cultural". In X Congreso Internacional da Asociación Internacional de Estudos Galegos (AIEG) 2012. Galiza alén do Arco Altántico. Unha visión multidisciplinar dos estudos galegos, Caerdydd/Cardiff, 12-14 september 2012.

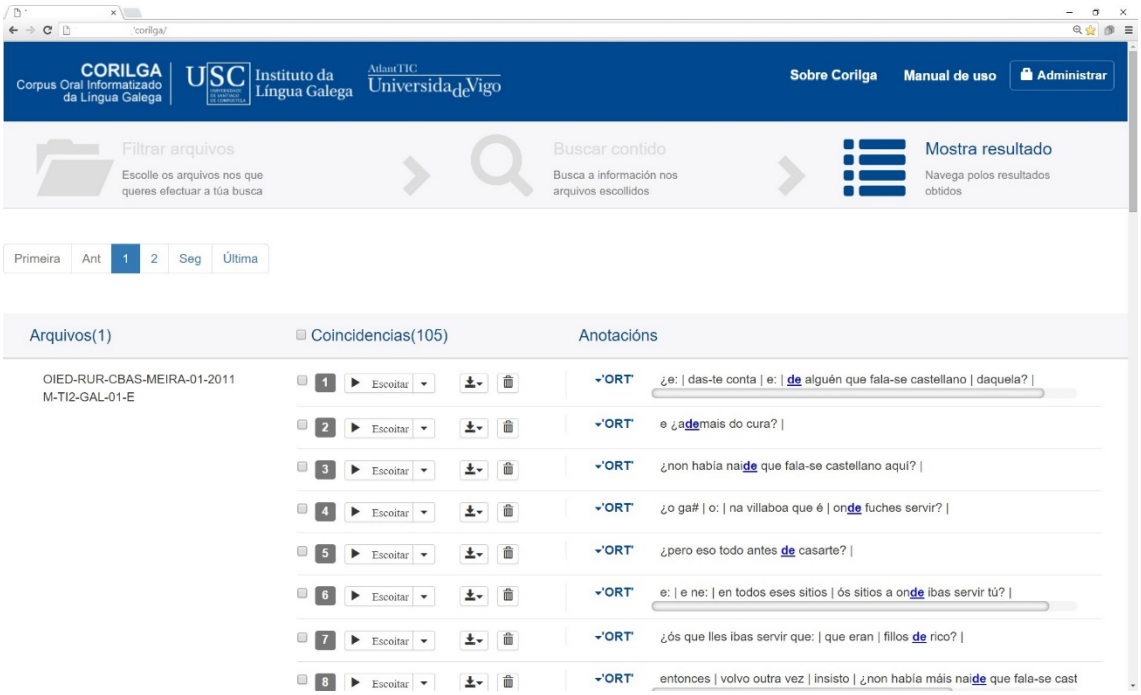

Figure 1: Example of the output for the search pattern “de”

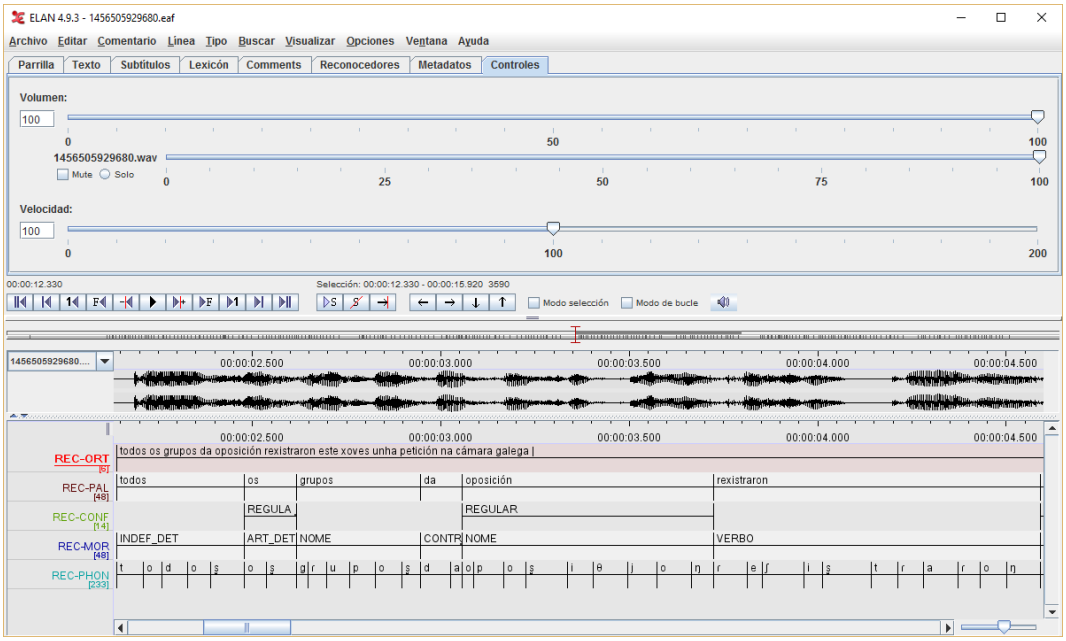

Figure 2: Example of the different levels of output for the recognition process

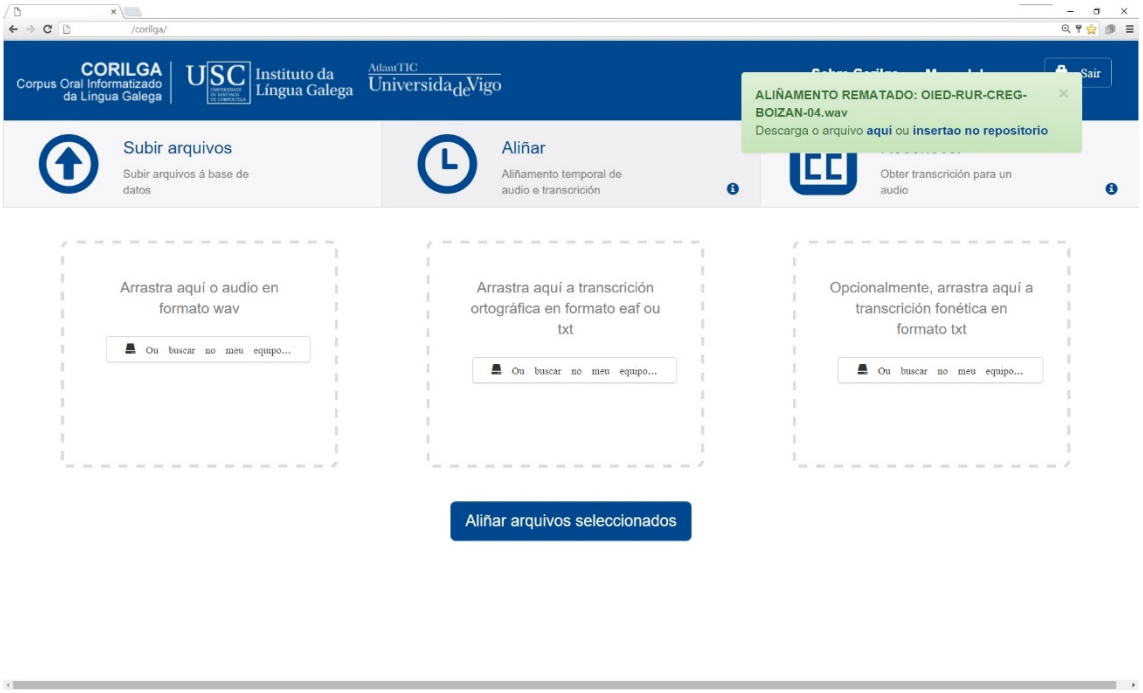

Figure 3: User interface for the alignment tool

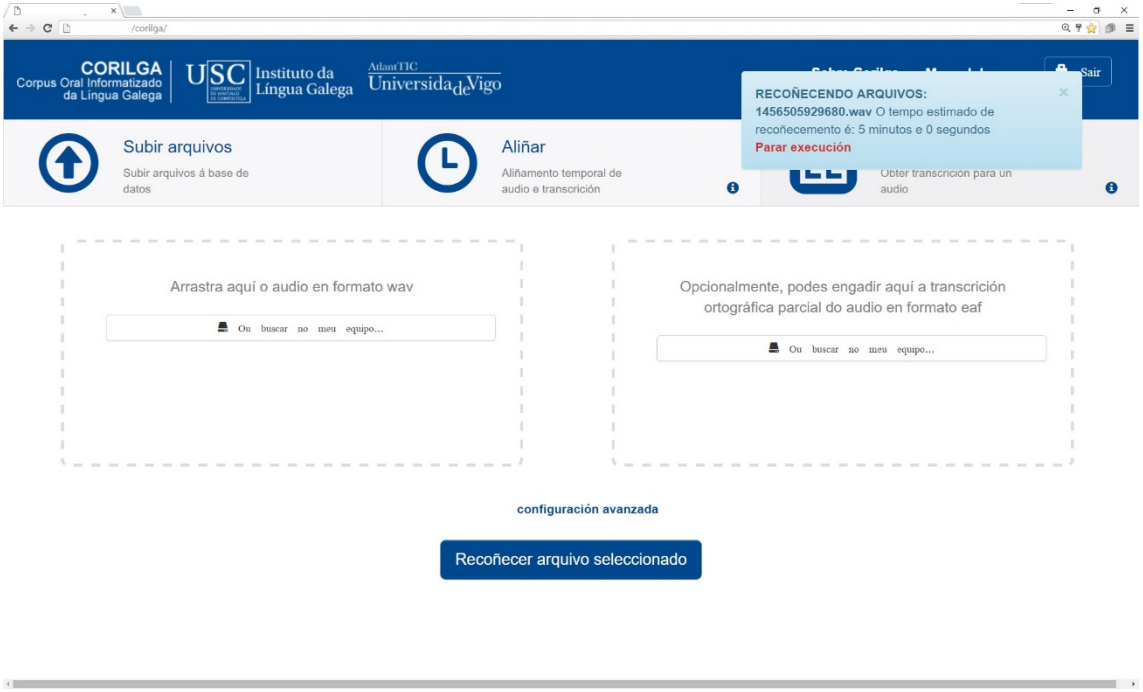

Figure 4: User interface for the recognition tool