The Nós Project: Opening routes for the Galician language in the field of language technologies

Conference Paper · June 2022

CITATION
1

READS
51

12 authors, including:

Iria De-Dios-Flores
University of Santiago de Compostela
9 PUBLICATIONS **14** CITATIONS
SEE PROFILE

Carmen Magariños
University of Vigo
11 PUBLICATIONS **93** CITATIONS
SEE PROFILE

Adina Vladu
University of Santiago de Compostela
12 PUBLICATIONS **1** CITATION
SEE PROFILE

John E. Ortega
New York University
37 PUBLICATIONS **271** CITATIONS
SEE PROFILE

Proceedings of the TDLE Workshop @LREC2022, pages 52–61
Marseille, 20 June 2022

The *Nós* Project: Opening routes for the Galician language in the field of language technologies

Iria de-Dios-Flores[1], Carmen Magariños[2], Adina Ioana Vladu[2], John E. Ortega[1], José Ramom Pichel[1], Marcos Garcia[1], Pablo Gamallo[1], Elisa Fernández Rei[2], Alberto Bugarín[1], Manuel González González[2], Senén Barro[1], Xosé Luis Regueira[2]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS), [2]Instituto da Lingua Galega (ILG)
Universidade de Santiago de Compostela
{iria.dedios, mariadelcarmen.magarinos, adina.vladu, john.ortega, jramom.pichel, marcos.garcia.gonzalez, pablo.gamallo, elisa.fernandez, alberto.bugarin.diz, manuel.gonzalez.gonzalez, senen.barro, xoseluis.regueira}@usc.gal

Abstract

The development of language technologies (LTs) such as machine translation, text analytics, and dialogue systems is essential in the current digital society, culture and economy. These LTs, widely supported in languages in high demand worldwide, such as English, are also necessary for smaller and less economically powerful languages, as they are a driving force in the democratization of the communities that use them due to their great social and cultural impact. As an example, dialogue systems allow us to communicate with machines in our own language; machine translation increases access to contents in different languages, thus facilitating intercultural relations; and text-to-speech and speech-to-text systems broaden different categories of users' access to technology. In the case of Galician (co-official language, together with Spanish, in the autonomous region of Galicia, located in northwestern Spain), incorporating the language into state-of-the-art AI applications can not only significantly favor its prestige (a decisive factor in language normalization), but also guarantee citizens' language rights, reduce social inequality, and narrow the digital divide. This is the main motivation behind the *Nós* Project (*Proxecto Nós*), which aims to have a significant contribution to the development of LTs in Galician (currently considered a low-resource language) by providing openly licensed resources, tools, and demonstrators in the area of intelligent technologies.

Keywords: Language technologies, Galician, linguistic rights, low-resource languages.

1. Introduction

Proxecto Nós[1] (The *Nós* Project) is an initiative promoted by the Galician Government (Xunta de Galicia), aimed at providing the Galician language (co-official language, together with Spanish, in the autonomous region of Galicia, located in northwestern Spain) with openly licensed resources, tools, demonstrators and use cases in the area of intelligent language technologies. The execution of *Proxecto Nós* has been entrusted to the University of Santiago de Compostela and is currently being carried out by a research team comprising members of the Instituto da Lingua Galega (ILG) and the Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS). *Nós* was planned as an ambitious initiative aimed to attract over €15 million in European funding. This paper presents the overall objectives taking this funding into account. Currently, the project is funded through a yearly agreement between the Galician Government and the University of Santiago de Compostela. This led to adopting a narrower focus for the Project's beginning (2021-2022), whose initial planning and documentation phase started in the last trimester of 2021 and ended in the first trimester of the current year.

The first stage of *Nós*, spanning from 2021 to 2025, will lay the foundations and provide the resources that will help place Galician among the languages that realize their full potential in the digital society and economy. The resources, tools and applications created within the *Nós* Project will improve technological support for Galician in order to achieve full digital language equality for both the present and the future. The ultimate goal is that Galician, as a low-to-medium resource language, will have reached the state where it has the necessary digital resources available to prosper as what is known as a "living" language in the digital age (Gaspari et al., 2021).

In an effort to generate the technological support and situational context necessary for Galician to continue to exist and prosper as a living language in the digital age, *Proxecto Nós* has been set up to address several areas from the natural language processing (NLP) realm. Specifically, the project is organized into several subprojects (jointly established by the Project's research team and Xunta de Galicia) where each sub-project corresponds to a major field from NLP. The eight subfields are the following: (i) speech synthesis, (ii) speech recognition, (iii) dialogue systems, (iv) error detection, (v) machine translation (MT), (vi) text generation, (vii) information extraction (IE), and (viii) opinion mining and fact checking. *Proxecto Nós* aims to address these eight sub-fields by first creating linguistic and computational resources in order to then build applications

[1]https://nos.gal/

based on these resources. Such applications will act as visible and accessible demonstrators of the technology developed in the Project and will, in turn, produce a tractor effect that will lead to the development of new products (use cases).

2. Background

2.1. Context and Motivation

The development of language technologies is a strategic innovation area geared towards augmenting the digital presence of a language in a society. It has been a priority in both Spanish (e.g. Plan Estatal de Investigación Científica y Técnica y de Innovación, Estrategia Española de Ciencia y Tecnología y de Innovación) and European scientific planning (e.g. Horizon 2020). Technologies such as MT, IE, text analytics, and conversational systems play a critical role in the digital society, culture, and economy.

Currently, linguistic data make up a very large part of the ever-increasing wealth of big data (Jill Evans et al., 2018). High-demand languages, such as English, benefit from a large amount of computational resources which can help to develop NLP software and tools. These languages benefit from a long-standing research tradition in different areas of language development, and their integration into artificial intelligence (AI) applications associated with the latest electronic devices such as conversational AI or automatic dictation software is high. Several language projects receive governmental funding such as the variety of projects financed by DARPA. Other languages that have come later to the table of AI research, such as Chinese, are currently following in the footsteps of English. Projects for Chinese, like the one from Qian Yan at Baidu, offer significant improvements for this language.

For those languages that cover a smaller set of speakers and, thus, are in lower demand, there exist efforts from local governments and other agencies to increase their use. A few projects similar to *Nós* include AINA, which aims to develop digital and linguistic resources for Catalan, several projects carried out at the HiTZ Research Center for Basque, CorCenCC for Welsh in Great Britain, and UQAILAUT for Inuktitut in Canada. The democratization of language technologies has a great social and cultural impact on the communities that use them (Ahmed and Wahed, 2020). For instance, MT increases access to contents in different languages, thus facilitating intercultural relations; dialogue systems allow us to communicate with machines in our own language; and semantic technologies enable advances in the automatic comprehension of texts, thus making it possible to process enormous quantities of documents. In the case of less-resourced languages such as Galician, the fact of incorporating the language into state-of-the-art AI applications can not only significantly favor its prestige (a decisive factor in language normalization), but also guarantee citizens' language rights, reduce social inequality, and narrow the digital divide (Jill Evans et al., 2018). Furthermore, the capacity to model language ensures a promising future for such technologies from both an economic and research and innovation perspective.

2.2. State of the Art: Galician Resources and Technologies

In 2012, work on Galician described a language with a level of technological support that "gives rise to cautious optimism". However, the authors also highlighted the need for creating new resources and tools for Galician, as well as directing more effort into LT (Language Technology) research, innovation, and development (García-Mateo and Arza, 2012).

In the last two decades, different research projects on Galician resulted in speech processing resources such as Cotovía(Rodríguez-Banga et al., 2012), the CORGA annotated reference corpus (Domínguez Noya et al., 2020) and other specialized corpora, both textual such as CLUVI (Gómez Guinovart, 2008), CTG (Gómez Guinovart, 2008), or TreeGal (García, 2016) and speech corpora like CORILGA (Regueira Fernández, 2012) and AGO (Rei, 2017). Furthermore, there are also functional morphosyntactic lemmatizers and taggers such as XIADA (Domínguez Noya, 2014), FreeLing (Gamallo and Garcia, 2013), and IXA-Pipes (Agerri et al., 2014), MT systems like GAIO (Xunta de Galicia, nd) or OpenTrad (Imaxin-Software, 2010), spellcheckers like OrtoGal (TALG, 2006 2019) and grammar checkers such as Avalingua (Gamallo et al., 2015). Also available are language analysis and information extraction tools like Linguakit (Gamallo et al., 2018) and language models such as SemantiGal (Garcia, 2021), Bertinho (Vilares et al., 2021), as well as other resources.

Furthermore, Galician is currently part of multilingual crowd-sourced data collection initiatives carried out by important companies on the global IT market, which have resulted in speech databases such as Google's SLR77 (Kjartansson et al., 2020) and Mozilla's Mozilla CommonVoice 7.0 and 8.0 (Ardila et al., 2020). This situation is reflected in a recent report on the current state of the LT field for Galician (Ramírez Sánchez and García Mateo, 2022), which informed on the considerable growth in the production of high-quality Galician resources and services, especially text resources. Despite the quality of these resources, it should be noted that not all are freely and publicly available for the development of LT.

The LT field has undergone profound changes over the last few years since the introduction of neural network systems. Generally, training models using these state-of-the-art technologies requires large quantities of data and has high energetic and computational costs, which continues to be a challenge for low-resource languages. However, as many recent studies show, end-to-end technologies and open-source multilingual pre-trained models created using large quantities of data from high-

resource languages (Shen et al., 2018; Baevski et al., 2020; Wolf et al., 2020) can be used, through transfer learning and fine-tuning, to train models in low-resource languages such as Catalan (Külebi and Öktem, 2018; Külebi et al., 2020) or, in our case, Galician. To this end, the existence of resources and tools that are freely available to the scientific and business community is essential, and that constitutes one of the main objectives of *Proxecto Nós*.

2.3. The potential of the Galician–Portuguese connection

One of the main reasons that the Galician language is excluded from several technological efforts is the lack of digital resources. As an under-resourced language (low-to-medium resources available), a large part of the effort that the *Nós* project will be dedicating its resources to is the compilation of corpora and other linguistic resources necessary for the development of computational models and algorithms. One advantage, however, that Galician has over other under-resourced languages is its close syntactic, semantic and orthographic similarity to Portuguese (Pichel et al., 2021), a high-resource language. Indeed, Galician and Portuguese are two closely-related members of the same language family (i.e. Galician-Portuguese).

In order to get an idea of the sheer amount of text resources available in Portuguese compared to Galician we can look at the amount of pages available for both languages on Wikipedia as an index of the online positioning of languages (a metric used by some companies like Google). There are more than a million pages in Portuguese while in Galician there are only 176.681 (Wikistats, 2022). Moreover, Portuguese not only has a large amount of text resources available but it also has a large scientific community involvement resulting in a large number of NLP tools.

The proximity between Galician and Portuguese is an advantage of incalculable value that puts us in a very good starting position, since the adaptation to Galician of most of the existing resources for Portuguese would be relatively simple. This is something that different researchers of the *Nós* project have been empirically demonstrating by using Portuguese to improve Galician resources (Garcia and Gamallo, 2010) and MT systems (Malvar et al., 2010). Furthermore, the close Galician-Portuguese relationship has led to members of the *Nós* project being among the organizing chairs of the PROPOR conference on the processing of the Portuguese language, where works on Galician can already be published, being considered as one variety of the Portuguese language space (Pinheiro et al., 2022).

3. Project Description

The *Nós* Project has two broad scientific and technological objectives: (i) to integrate the Galician language into cutting-edge AI and language technologies, thus enabling its use in human-machine interactions; and (ii) to produce a qualitative leap forward in the development of language technologies for Galician. For this purpose, resources, tools, and applications will be developed and distributed under open licenses, which will allow them to be integrated into existing devices and services (such as smart speakers or conversational agents) and future technologies.

To this end, specific objectives directly related to some of the major NLP tasks have been established. Each of these technological objectives will be executed as separate sub-projects which will allow the parallel development of different tasks and an overall more effective organization. Nonetheless, a set of general objectives are shared by all the tasks. The general objectives are: (i) the compilation of high-quality linguistic resources; (ii) the elaboration of language and acoustic models (both general-purpose and task-specific models); and (iii) the development of applications based on these models. In addition, the project will have a general coordination mechanism through which resources will be distributed and shared among the different subprojects.

The resources and models developed for each task will be made available to the public using common dissemination repositories (e.g. GitHub, Hugging Face) and platforms (e.g. European Language Grid) , thus allowing their use in all kinds of applications, services, and products, by the scientific community, companies, institutions, and society in general. The results will be disseminated through a repository available at the project's web portal (which can be hosted on internal servers), as well as other established and internationally recognized repositories. Finally, the project contemplates the complete development of applications based on these resources which will act as visible and accessible demonstrators of the developed technology and will produce a tractor effect that will lead to the creation of new products.

The general objectives, sub-projects and coordination strategy are further detailed in sections 4, 5 and 6, respectively.

4. General Objectives

The main objectives are described in further detail below:

- **Compilation and creation of linguistic resources.** In order to place the Galician language on equal terms with other languages in the digital sphere, it is an essential requirement to have a wide variety and large number of high-quality language resources (annotated reference corpora, web-scale corpora, task- and domain-specific corpora, parallel corpora, knowledge bases, dictionaries, etc.) that allow the development of cutting-edge technologies for Galician. These resources will be mainly created from zero, depending on the needs of each task. In addition, all the generated resources will be distributed under free

licences to encourage their extension, improvement, and exploitation by third parties.

- **Elaboration of language and acoustic models.** Besides providing the required linguistic resources, statistical and computational models will also be developed based on these resources, using different leading-edge techniques. Thus, both pre-trained general purpose models and models adapted to specific tasks and domains will be developed by applying state-of-the-art techniques, mainly neural network-based deep learning. As with the linguistic resources, these models will be made publicly available under free licences, allowing them to be freely used by companies, institutions and end users.

- **Development of applications.** Lastly, as final demonstrators of the project, a set of fully functional applications will be developed, which will showcase the potential of the elaborated models and resources. At this point, the project aims to build both specific demonstrators for each of the previously listed tasks, and applications that connect and integrate different technologies. These demonstrators, apart from illustrating the work carried out in the project, will also serve to foster the development of new products bringing together new technologies and the Galician language.

5. Subprojects

In what follows, we provide a summary of the specific objectives for each of the eight subprojects, including some brief technical details on the necessary linguistic and computational resources, the proposed demonstrators and possible use cases that could be developed by third parties. Given that the project is still in its initial stages, adjustments are expected to take place along its execution during the next years.

5.1. Speech Synthesis

The objective of this subproject is to provide the necessary means (resources and technologies) so that intelligent devices and systems can speak in Galician, as a first step towards an interaction with these devices in Galician at the same level as other languages such as English or Spanish. To this end, we will create public datasets that allow the development of state-of-the-art text-to-speech conversion systems, with the ability to produce synthetic speech with different identities, styles and emotions. These datasets will contain voice recordings obtained from phonetically balanced corpora, with their corresponding textual transcription. In addition to being distributed publicly, the data generated will be used to train different voice models, both speaker-dependent and average voice models (AVM).
As demonstrators of this subproject, the following applications could be developed:

- **High quality text-to-speech (TTS) conversion system**, with the ability to produce synthetic speech with different speaker identities (possibility to choose gender, age, etc.), speaking styles (radio news, conversation, advertising speech, etc.) and emotional expressions (sadness, surprise, joy, etc.). Among the use cases of the text-to-speech conversion system, in addition to its integration in general-purpose applications (virtual assistants, dialogue systems, automatic translators, etc.), it is worth mentioning the possibility of incorporating it into applications for people with disabilities. For instance, screen readers and audio description systems for visually impaired people, or mobile TTS applications specially designed for people with speech disorders or impairments.

- **Web interface for obtaining personalized synthetic voices**, which would allow end-users to obtain a personalized speech synthesizer with their own voice. To obtain these personalized voices, users will have to record a small number of sentences that will be used to adapt a pre-trained AVM. The objective of this demonstrator would be to show the potential in terms of adaptability of the AVMs, trained from several voices. As possible use cases of this web interface, besides the possibility of integrating such voices in different devices (e.g. GPS), this interface could be used to create voice backups (Erro et al., 2015; Erro et al., 2014). This latter development would be particularly useful for people who suffer from some kind of pathology (e.g. people who have to undergo a surgery that involves the loss of voice) to preserve the ability to communicate with their own voice. This interface would also allow applying cross-lingual adaptation techniques (Magariños et al., 2019), with the aim of obtaining personalized synthetic voices in a different language from the user's original language. More specifically, these techniques would allow to obtain customized synthesizers with the user's voice in languages other than Galician (e.g. English, Spanish or Portuguese). As a direct application, these synthesizers would allow to customize speech-to-speech translation systems, so that the voice identity of the original speaker can be preserved in the translated speech.

5.2. Speech Recognition

Together with the previous subproject, the ultimate goal of the speech recognition subproject is to enable a complete oral interaction in Galician between users and intelligent devices. In order to achieve this goal, it will be necessary to generate and distribute publicly a large set of speech and text corpora, needed to train acoustic and language models, respectively. The proposal of potential demonstrators for this task is as follows:

- **General purpose automatic speech recognition (ASR) system**, able to perform across different domains. As use cases, one of the main applications of an ASR system is to provide a voice user interface for other systems such as virtual assistants, dialogue systems, web browsers or automatic translation systems. Moreover, one should not forget the possibilities offered by ASR systems in terms of voice commands for intelligent home devices (lighting control, thermostats, intruder alarms and all kinds of household appliances).

- **Automatic subtitling system.** As a use case, this system could be used to develop a real-time automatic subtitling tool for Galician newscasts.

- **Personalized automatic dictation system.** The possible use cases of this demonstrator are its integration in office software (document writing) or mail servers (e-mail writing). Another interesting application would be note-taking during medical consultations. This system would allow capturing patient diagnosis notes automatically, reducing the average duration of consultations.

5.3. Dialogue Systems

The main objective of this subproject is to provide guides and packages of specific technological and linguistic resources that facilitate the construction of competent conversational agents in Galician. In order for conversational agents to be linguistically and socially competent, they must be provided with additional mechanisms that allow them to handle conversational contexts that go beyond the specific task or scope of use of the conversation. In addition to a refinement of conversation tracking techniques, we will combine machine learning with knowledge representation (e.g. semantic networks and graphs) to reduce the amount of linguistic data needed and take advantage of existing resources from other languages. This is especially important in the context of Galician, which lacks a large annotated corpus to be used in dialogue systems.
As demonstrator of this subproject, we plan to develop the following application:

- **Chatbot-like conversational agent** to focus the interaction on the input and output of text in Galician. Among the most interesting use cases we highlight the specialization of the conversational agent for task-oriented application domains, for example the management of citizen appointments for the public administration, and for a more general scope (tourism, integration of cultural information, etc.).

5.4. Automatic Error Detection

The linguistic correction and evaluation subproject aims to provide the Galician language with a series of improved applications which make it possible to verify, correct and evaluate texts automatically at different levels using natural language processing techniques to detect and classify errors or deviations. To achieve this goal, we will design computer programs for spelling, grammar and style correction based on already existing tools adapted to Galician (e.g. Galgo by imaxin|software and Xunta de Galicia or AvalinguaCiTIUS). These tools will be improved with the most advanced techniques so that the identification of errors regarding words in context (spelling correction) and sequences of words and structures (grammatical correction) reaches the state of the art for major languages. Thus, we will not limit ourselves to the identification of errors, but also on the use of appropriate or preferred linguistic structures, lexical-semantic content and density, and the coherence and fluency of a text. For the latter, it is necessary to have libraries and linguistic and computational resources of greater complexity, which can represent the semantic content by grouping words and categorize texts using statistical methods or through automatic learning based on texts labeled by expert evaluators. As possible demonstrators of this subproject, we plan to develop the following online applications (which would be integrated in a single tool):

- **Spelling, grammar and style proofreader of texts.**

- **Linguistic quality evaluator.**

- **Tone and sensitivity analyzer.**

These applications would provide the Galician language with a series of valuable resources to improve the quality of the written language in all areas (education, press, private sector, etc.). In addition, they could be added to different programs or contexts through use cases adapted by third parties (browsers, email managers, office software, intelligent keyboards in mobile phones/tablets, etc.). For example, as a domain-specific use case, they could be adapted for the automatic evaluation of the linguistic quality of the works elaborated by high school and university students, of the entries in Galipedia, or even in order to adapt past literary resources to the current Galician standard. Another socially relevant use case is a correction/management system that promotes inclusive language.

5.5. Machine Translation

The MT subproject consists of the development of neural translation systems that allow both native speakers of Galician and professional companies and institutions to translate texts and short documents quickly and accurately. At present, there are different automatic translation systems with a linguistic rule-based approach, such as the open source automatic translation service platform Opentrad of the company imaxin—software, which is implemented and improved in the automatic

translator Gaio of Xunta de Galicia (AMTEGA). However, there are no state-of-the-art models based on neural AI techniques with a web interface for high quality translations in Galician.

Our proposal includes the development of the whole End-to-End system that consists of the research, implementation and testing of an on-line MT system that will compete with other MT systems of similar strategies as those provided by companies, such as Google, Microsoft, and Yandex, where Galician is offered as a source language to be translated, but the quality of translation is not yet at the level of the languages with more resources. The improvement of translation will be achieved by taking advantage of the highest quality models that will be created within the *Nós* project.

The proposal of demonstrators for this subproject is as follows:

- **NMT translator from Galician to other languages and vice versa.** The defined strategic language pairs are Galician-Spanish, Galician-Portuguese and Galician-English, although other pairs could be incorporated later on. As use cases, these general translating systems, including the linguistic resources used for their implementation, could be adapted for specific domains by third parties.

5.6. Text Generation

The main objective of this subproject is to focus on the development of computational and linguistic resources for the automatic natural language generation (NLG) of texts in Galician language. The focus is on data-to-text resources, since most of the resources of interest in the text-to-text area are already dealt with in other subprojects. We will address the two existing approaches for this type of systems: the traditional template-based approach and more recent end-to-end approaches based on different deep learning artificial neural network architectures. Both approaches present major scientific and technological challenges and require the development of computational and linguistic resources for the construction of impactful systems and applications. One of these challenges is the reliability of validation by automatic metrics of neural end-to-end systems, especially in critical applications. Thus, it will be necessary to design strategies to contrast and verify the texts generated by data-to-text neural systems with respect to the original data and to develop the necessary technology that facilitates this verification in a semi-automated way. A critical aspect when defining demonstrators in the NLG environment is their ability to transmit information to visually impaired people, when combined with text-to-speech systems, as well as to overcome the limitations of small display devices (mobiles, tablets, etc.) where graphical visualization is not suitable.

The demonstrators that we aim for in this area are:

- **Automatic generator of different types of visualization graphs** (time series, bar charts, trend charts, etc.) of a generic type, which are commonly used in all types of reports.
- **Automatic real-time data report generator**, with direct application in the industrial sector (ICT, production or industrial plants, logistics, etc.).
- **Abstractive summarizer** of a general type based on end-to-end models.

The technology developed for these demonstrators can prompt impactful use cases such as the automatic generator of weather forecasts and environmental information, the automatic generator of medical reports from the data available in the medical history, an automatic generator of banking or personal economy reports or an automatic generator of informative chronicles, among many others.

5.7. Information Extraction

This subproject will focus on developing a set of text-to-data techniques for discovering and extracting relevant and salient knowledge from large amounts of unstructured Galician text. Its main goal is to extract knowledge elements or regular patterns from a collection of documents, especially content extracted from the web and social media. Since the extraction task is a very broad set of techniques, before going into the description of the specific demonstrators, it is necessary to define several empirical challenges such as (i) semantic relation extraction, (ii) named entity recognition (NER), (iii) entity linking, (iv) event detection, (v) topic modelling for unsupervised document classification, or (vi) keyword and terminology extraction. Once these information extraction techniques have been dealt with, some examples of possible demonstrators that we plan to develop are:

- **Question answering system (QA) of encyclopedic character**. A QA system mainly based on unsupervised learning needs NER and relation extraction techniques to structure the information from a source corpus (e.g. Galician Wikipedia), and thus be able to map the question with the extracted information and candidate to be the most successful answer. The development of this system also involves the semi-automatic construction of a knowledge base or ontology where the extracted information is organized.
- **Semantic annotating tool with linked data** for the creation of teaching materials in Galician: By means of NER and entity linking, we proceed to the enrichment of teaching materials in Galician. For example, new information can be added to the terms and entities identified in a text by linking them to the corresponding entries in the Galician Wikipedia.

- **Document classifier.** This system can be applied to any collection of Galician documents to be organized and grouped into non-predefined clusters. For this demonstrator, it will be necessary to take into account terminology extraction and topic modeling algorithms.
- **Extractive summarizer** that extracts relevant chunks from input documents to build summaries of different sizes, according to the users' needs. Extractive summarization techniques based on prior identification of relevant keywords and multiwords will be explored.

In addition to these demonstrators, the tools and models elaborated have the potential of giving rise to a wide variety of use cases from third parties, such as a process model generator from medical reports or a detector of depression signals in social networks, which would require most of the information extraction techniques developed in the scope of the project.

5.8. Opinion Mining and Fact Checking

This subproject will focus on providing the necessary resources and technologies to carry out basic opinion mining and fact checking analyses in Galician. To carry out the task of opinion mining, we will explore unsupervised strategies based on the use of polarized lexicons as well as supervised techniques dependent on annotated corpora with information about the polarity of the texts. Besides the generic models created with large corpora, it will be necessary to elaborate polarized lexicons using automatic and semi-automatic techniques, as well as to annotate new datasets with polarized texts. The classifiers created will take into account morphosyntactic and syntactic analysis to deal more correctly with complex linguistic phenomena such as negation and compositionality (Vilares et al., 2017). They will also take into account the information provided by a NER for the identification of entities mentioned in the text. This information is crucial in order to implement surveillance and monitoring systems on those entities, namely, products, companies, organizations, people, etc.

Regarding the verification or checking of factual information, we will distinguish, on the one hand, the process of extracting verifiable factual information (or facts) and, on the other hand, the process of verification of the factual content itself. For the first process, it is essential to consider factual information extraction tools, such as those derived from the extraction of open information, focused on the identification of basic propositions in the input text. It is also essential to have information resources where news and information contrasted in reliable sources and knowledge bases are compiled. For the checking process, we use textual semantic techniques focused on the computation of sentence similarity with neural networks and transformers. These techniques allow us to compare the sentences that convey factual information, extracted from the input texts, with truthful data previously contrasted in knowledge bases and reliable sources.

Taking the above into account, we consider building the following demonstrators:

- **Monitoring system of Galician products, companies and organizations.**
- **Map of the best-valued Galician locations in real time.**
- **Bilingual (Galician-Spanish) news checker.**
- **App focused on the identification of Galician toxic bots on Twitter.**

From the generic linguistic resources and language models used to implement these demonstrators, it will be possible to successfully develop use cases adapted to specific domains, such as, for example, a system for monitoring the products of a specific company, a map of top-rated Galician tourist sites, or a fact checker in the health domain specialized in detecting false rumors.

6. Project Management

The established scientific and technological objectives are organized into a set of work packages (WPs) whose interrelation is illustrated in Figure 1. Each of these WPs, briefly described below, comprises a series of tasks and deliverables that guarantee the fulfillment of the project's objectives:

- **WP1 – Global project management.** The goal of this work package is to ensure the effective coordination and management of the different subprojects as well as the overseeing of the deliverables, without forgetting the ethical and legal dimensions.
- **WP2 – Monitoring, evaluation, and continuous contribution to the state of the art of science and technology.** This work package comprises the development of a methodology for monitoring the state of the art of all the tasks involved in the project. This will allow the design of an experimentation plan suitable for each of the subprojects, which will be developed in the work packages 3, 4 and 5. This design might be updated as the state of the art evolves.
- **WP3 – Obtaining and creating high-quality language resources.** The aim of this work package is to develop different types of high quality corpora (spoken and written) containing deep-annotated linguistic information. In general, three types of corpora will be needed: on the one hand, a reference corpus for Galician and a web macrocorpus, both to be used by all the subprojects; and, on the other hand, different purpose-specific corpora (datasets).

- **WP4 – Construction and evaluation of state-of-the-art language and acoustic models.** This package will focus on the development of different models for all the subprojects and tasks (language models, acoustic models, etc.) from the resources obtained in package 3. This will involve the use of different linguistic tools (either adapted or created from zero) and state-of-the-art deep learning techniques.
- **WP5 – Development and evaluation of demonstrators and use cases.** The purpose of this package is the development of the agreed set of demonstrators and use cases, as well as the methodologies and systems necessary for their proper evaluation, whether perceptual or automatic. To this end, both objective and perceptual performance measures will be used.
- **WP6 – Publication and dissemination.** This package includes the publication of the project's outcomes (models, tools, demonstrators and use cases) for their general use, the divulgation of scientific results (publication in specialized journals and conference proceedings), as well as the launch of different calls for interest. It also envisages the development of an online repository that will allow not only testing the different demonstrators, but also freely downloading all the resources and the associated source code.

The central workflow of the project encompasses work packages 3, 4 and 5. These packages are the cornerstone of the project since they focus on the most relevant scientific and technological tasks: the development of linguistic resources, language and acoustic models and demonstrators.

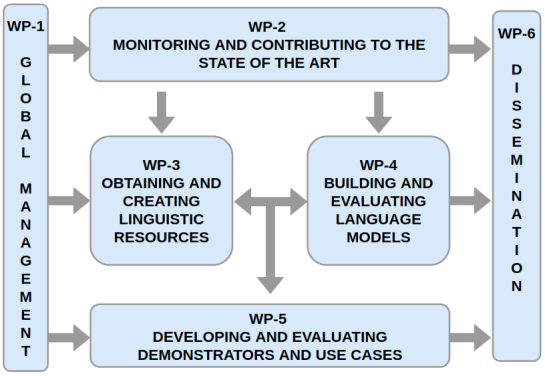

Figure 1: Interrelation among the different work packages that make up the project.

7. Current Developments and Future Work

Among the initial results of the project, we can highlight the first crawl of a web-based Galician corpus and a language model based on the CCNet tools and data (Ortega et al., 2022b), the development and testing of two BERT language models (with 12 and 6 layers, respectively) (Garcia, 2021), as well as the development and testing of a Spanish-Galician neural machine translation (NMT) system prototype (Ortega et al., 2022a). For the current year, *Proxecto Nós* aims to keep generating linguistic and computational resources to explore different subprojects. Specifically, work will be carried out on the design and recording of a high-quality speech corpus of sufficient size so as to allow the training of TTS state-of-the-art models. On the other hand, a speech corpus for ASR will be compiled. In addition, parallel Galician-Spanish, Galician-English, and Galician-Portuguese corpora will be compiled and used, together with existing multilingual corpora, for the development of NMT systems. Additionally, a web-scale Galician text corpus will be compiled, larger than the one already constructed, to be used in all the subprojects working with written text included in Nós. Based on these resources, new language models will be developed using different state-of-the-art techniques, as well as demonstrators or prototypes of a TTS system, translation system, and automatic text generator for Galician. At the same time, efforts will focus on extending and improving the first systems developed, and on validating the results obtained via the creation of high-quality gold standards.

8. Acknowledgements

This research was funded by the project "Nós: Galician in the society and economy of artificial intelligence" (Proxecto Nós: O galego na sociedade e economía da intelixencia artificial 2021-CP080), agreement between Xunta de Galicia and University of Santiago de Compostela, and grant ED431G2019/04 by the Galician Ministry of Education, University and Professional Training, and the European Regional Development Fund (ERDF/FEDER program).

9. Bibliographical References

Agerri, R., Bermudez, J., and Rigau, G. (2014). Ixa pipeline: Efficient and ready to use multilingual nlp tools. In *Proceedings of the 9th Language Resources and Evaluation Conference (LREC2014)*, pages 26–31.

Ahmed, N. M. and Wahed, M. (2020). The de-democratization of ai: Deep learning and the compute divide in artificial intelligence research. *ArXiv*, abs/2010.15581.

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., and Weber, G. (2020). Common voice: A massively-multilingual speech corpus. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 4211–4215.

Domínguez Noya, E. M., López Martínez, M. S., and Barcala Rodrígez, F. M. (2020). *O Corpus*

de Referencia do Galego Actual (CORGA): composición, codificación, etiquetaxe e explotación. Corpus y construcciones. Perspectivas hispánicas. Marta Blanco, Hella Olbertz and Victoria Vázquez Rozas (Series Editors). Universidade de Santiago de Compostela.

Domínguez Noya, E. M. (2014). Etiquetación y desambiguación automáticas en gallego: el sistema XIADA. 52:93–96.

Erro, D., Hernáez, I., Navas, E., Alonso, A., Arzelus, H., Jauk, I., Hy, N. Q., Magarinos, C., Pérez-Ramón, R., Sulır, M., et al. (2014). Zuretts: Online platform for obtaining personalized synthetic voices. *Proc. eNTERFACE*, 14.

Erro, D., Hernaez, I., Alonso, A., García-Lorenzo, D., Navas, E., Ye, J., Arzelus, H., Jauk, I., Hy, N. Q., Magariños, C., et al. (2015). Personalized synthetic voices for speaking impaired: Website and app. In *Sixteenth Annual Conference of the International Speech Communication Association*.

Gamallo, P. and Garcia, M. (2013). Freeling e treetagger: um estudo comparativo no âmbito do português. *Relatório técnico. Universidade de Santiago de Compostela*.

Gamallo, P., Garcia, M., del Río, I., and González López, I. (2015). *Avalingua: Natural language processing for automatic error detection*. John Benjamins.

Gamallo, P., García, M., Piñeiro, C., Martinez-Castaño, R., and Pichel, J. C. (2018). LinguaKit: A Big Data-Based Multilingual Tool for Linguistic Analysis and Information Extraction. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 239–244.

Garcia, M. and Gamallo, P. (2010). Análise morfossintáctica para português europeu e galego: Problemas, soluções e avaliação. *Linguamática*, 2(2):59–67, Mai.

Garcia, M. (2021). Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL-IJCNLP)*, pages 3625–3640.

García-Mateo, C. and Arza, M. (2012). *O idioma galego na era dixital – The Galician Language in the Digital Age*. META-NET White Paper Series. Georg Rehm and Hans Uszkoreit (Series Editors). Springer.

García, M. (2016). Universal dependencies guidelines for the galician-treegal treebank. technical report. Technical report, LyS Group, Universidade da Coruña.

Gaspari, F., Way, A., Dunne, J., Rehm, G., Piperidis, S., and Giagkou, M. (2021). Digital language equality (preliminary definition). Technical report, European Language Equality (ELE) Consortium.

Gómez Guinovart, X. (2008). *A investigación en lexicografía e terminoloxía no Corpus Lingüístico da Universidade de Vigo (CLUVI) e no Corpus Técnico do Galego (CTG)*. A lexicografía galega moderna. Recursos e perspectivas. González Seoane, Ernesto, Antón Santamarina and Xavier Varela Barreiro (eds.). Santiago de Compostela: Consello da Cultura Galega / Instituto da Lingua Galega.

Imaxin-Software. (2010). Opentrad: machine translation.

Jill Evans, rapporteur, C. o. C., Education (CULT), Committee on Industry, R., and (ITRE), E. (2018). Report on language equality in the digital age. Technical report, European Parliament, Strasbourg, France.

Kjartansson, O., Gutkin, A., Butryna, A., Demirsahin, I., and Rivera, C. (2020). Open-source high quality speech datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 21–27, Marseille, France, May. European Language Resources association.

Külebi, B. and Öktem, A. (2018). Building an Open Source Automatic Speech Recognition System for Catalan. In *Proc. IberSPEECH 2018*, pages 25–29.

Külebi, B., Öktem, A., Peiró-Lilja, A., Pascual, S., and Farrús, M. (2020). CATOTRON — A Neural Text-to-Speech System in Catalan. In *Proc. Interspeech 2020*, pages 490–491.

Magariños, C., Erro, D., and Banga, E. R. (2019). Language-independent acoustic cloning of hts voices. *Computer Speech Language*, 55:168–186.

Malvar, P., Pichel, J. R., Senra, , Gamallo, P., and García, A. (2010). Vencendo a escassez de recursos computacionais. carvalho: Tradutor automático estatístico inglês-galego a partir do corpus paralelo europarl inglês-português. *Linguamática*, 2(2):31–38, Mai.

Ortega, J. E., de-Dios-Flores, I., Campos, J. R. P., and Gamallo, P. (2022a). A neural machine translation system for spanish to galician through portuguese transliteration. Demo presented at the 15th International Conference on Computational Processing of Portuguese (PROPOR 2022).

Ortega, J. E., de-Dios-Flores, I., Campos, J. R. P., and Gamallo, P. (2022b). Revisiting ccnet for quality measurements in galician. In Vládia Pinheiro, et al., editors, *Computational Processing of the Portuguese Language - 15th International Conference, PROPOR 2022, Fortaleza, Brazil, March 21-23, 2022, Proceedings*, volume 13208 of *Lecture Notes in Computer Science*, pages 407–412. Springer.

Pichel, J. R., Gamallo, P., Alegria, I., and Neves, M. (2021). A methodology to measure the diachronic language distance between three languages based on perplexity. *Journal of Quantitative Linguistics*, 28(4):306–336.

Vládia Pinheiro, et al., editors. (2022). *Computational*

Processing of the Portuguese Language: 15th International Conference, PROPOR 2022. Lecture Notes in Computer Science, Springer.

Ramírez Sánchez, J. and García Mateo, C. (2022). Report on the galician language (deliverable d1.15). Technical report, European Language Equality.

Regueira Fernández, X. L. (2012). *Corpus Oral Informatizado da Lingua Galega*. Santiago de Compostela, Universidade de Santiago.

Rei, F. F. (2017). O arquivo do galego oral: xénese e situación actual. In *Gallæcia: Estudos de lingüística portuguesa e galega*, pages 545–564. Universidade de Santiago de Compostela.

Rodríguez-Banga, E., García-Mateo, C., Méndez-Pazó, F., González-González, M., and Magariños, C. (2012). Cotovía: an open source TTS for Galician and Spanish. In *Proc. IberSPEECH 2012: VII Jornadas en Tecnología del Habla and III Iberian SLTech Workshop*, pages 308–315. RTTH and SIG-IL.

TALG, S. (2006-2019). Ortogal.

Vilares, D., Gómez-Rodríguez, C., and Alonso, M. A. (2017). Universal, unsupervised (rule-based), uncovered sentiment analysis. *Knowledge-Based Systems*, 118:45–55.

Vilares, D., Garcia, M., and Gómez-Rodríguez, C. (2021). Bertinho: Galician BERT Representations. *Procesamiento del Lenguaje Natural*, 66:13–26.

Wikistats. (2022). Anexo:wikipedias. *Wikipedia-Wikimedia Foundation*. Available online at `https://es.wikipedia.org/wiki/Anexo:Wikipedias`.

Xunta de Galicia, S. X. d. P. L. (n.d.). Gaio: Tradutor automático.