EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

BEST-Assistants – Building E2E Spoken-Language Understanding Systems for Virtual Assistants in Low-Resources Scenarios

Authors	Andrés Piñeiro Martín, María del Carmen López Pérez, Carmen García Mateo (UVIGO), Laura Docío Fernández (UVIGO)
Organisation	Balidea, University of Vigo
Dissemination level	Public
Date	31-03-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	BEST-Assistants – Building E2E Spoken-Language Understanding Systems for Virtual Assistants in Low-Resources Scenarios
Authors	Andrés Piñeiro Martín, María del Carmen López Pérez, Carmen García Mateo (UVIGO), Laura Docío Fernández (UVIGO)
Organisation	Balidea, University of Vigo
Type	Report
Number of pages	42
Status and version	Final
Dissemination level	Public
Date of delivery	31-03-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2023 ELE2 Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l'égalité des langues (European Language Equality Network)	ELEN	FR

Contents

1 Introduction **1**

2 Motivation: Creating and obtaining a dataset for SLU from scratch **3**
2.1 Task definition . . . 3
2.2 The idea of the project . . . 4
2.3 FalAI voice collection tool . . . 5

3 Dataset design **9**
3.1 Benchmark SLU datasets . . . 10
3.2 Dataset complexity . . . 10
3.2.1 Lexical analysis: n-Gram Entropy . . . 11
3.2.2 Syntactic analysis: lexical sophistication (LS2) . . . 11
3.2.3 Semantic analysis: semantic textual similarity (STS) . . . 12
3.3 FalAI dataset features . . . 13
3.3.1 Minimum design features . . . 13

4 Campaigns design **14**
4.1 Media campaign . . . 16
4.2 Young people campaign . . . 17
4.3 Older people campaign . . . 19

5 FalAI evolution **19**
5.1 Official launch . . . 20
5.2 Youth Campaign Launch . . . 20
5.3 First reactions and campaign reception . . . 20
5.4 Midterm analysis and older people campaign evolution . . . 22
5.5 FalAI results . . . 22

6 Validation and analysis over the collected dataset **25**
6.1 Semi-automatic validation . . . 26
6.2 Quality metrics over collected dataset . . . 27

7 Lessons learned **28**
7.1 Key points in the approach . . . 28
7.2 Key points for participation . . . 30
7.3 Actions where measuring the impact has not been possible . . . 33

8 Conclusions **34**
8.1 Future directions . . . 34

List of Figures

1 Initial screen and information entry screen of falAI. 6
2 Recording screen of falAI. 7
3 STS between domains in falAI calculated with LaBSE. 12
4 STS between intents and sentences in falAI calculated with LaBSE. 12
5 Objectives in terms of number of recordings and number of hours. 15
6 Galicia map with distribution of participants and municipalities with the highest number of participants. 15
7 Number of participants per province and the distribution according to gender. 16
8 Distribution of participants according to accent and a motivational graph with the number of hours achieved. 16
9 Screen to enter the draw after completing the 30 sentences. 17
10 Survey and participant data for the draw. 18
11 Evolution of falAI: number of recordings per day during the campaign. 19
12 Message designed to broadcast via instant messaging. 21
13 Number of recordings per hour after tweet viralisation. 22
14 Participants per municipality. 23
15 Participants per province. 24
16 Participants per accent. 25
17 Participants per age range. 26
18 How people have known falAI. 30
19 Number of recordings per hour after the collaboration of two influencers. . . . 31
20 Reasons for participation in falAI provided by users. 32

List of Tables

1 Text corpora SLU dataset comparison. 10
2 Comparison of entropies between the main SLU datasets and falAI dataset. . . 11
3 KPI objectives estimation for campaigns. 15
4 FalAI main results. 23
5 FalAI province results in terms of hours and recordings. 24
6 FalAI accent results. 25
7 FalAI age range results. 25
8 Metadata labelling in manual validation. 27
9 Spoken dataset comparison. 28
10 Quality measurements over collected dataset. 28

List of Acronyms

AI	Artificial Intelligence
ASR	Automatic Speech Recognition
ATIS	Air Travel Information System
CER	Character Error Rate
DoS	Denial of Service
EC	European Commission
E2E	End-to-End
ELE	European Language Equality

ELE2	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
ELT	European Language Technology
FSC	Fluent Speech Commands
GDPR	General Data Protection Regulation
GTM	Multimedia Technologies Group
KPI	Key Performance Indicator
LLM	Large Language Model
LS	Lexical sophistication
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
MUVI	Galician Videogame Museum
NCC	National Competence Centre
NLG	Natural Language Generation
NCP	National Contact Point
NLP	Natural Language Processing
NLU	Natural Language Understanding
SRIA	The Strategic Research and Implementation Agenda
SLU	Spoken Language Understanding
SLURP	Spoken Language Understanding Resource Package
STOP	Spoken Task Oriented semantic Parsing
STS	Semantic Textual Similarity
STOA	Science and Technology Options Assessment
STT	Speech To Text
UVIGO	University of Vigo
VA	Virtual Assistant
WER	Word Error Rate
XLS-R	Cross-Lingual Speech Representation

Abstract

A spoken language understanding (SLU) system is traditionally designed as a pipeline with an ASR (automatic speech recognition) module followed by an NLU (natural language understanding) module, where each component is optimised independently. Creating more efficient end-to-end (E2E) SLU architectures is a hot topic in research due to the growing demand for speech interfaces. However, these systems require huge amounts of speech data for training (not available and expensive to obtain) and they are difficult to adapt to new domains (specific data has to be generated). This report presents guidance for designing and obtaining datasets for E2E SLU systems in low-resources scenarios. The use case will be a virtual assistant applied in telemedicine for chronic patients.

1 Introduction

In the last decade, there has been a increasing interest by the technology companies, but also by governments and administrations, in the development of Virtual Assistants (VAs). The latest breakthroughs in fields such as Natural Language Understanding (NLU), Automatic Speech Recognition (ASR) or Large Language Models (LLM) make it possible to communicate with machines in a more natural and fluent way, in broader contexts, normalizing voice-based interactions with the virtual world. Virtual assistants have the potential to become the main channel of communication and information extraction with the digital world, and they are expected to continue to grow exponentially over the next decade. Assistance robots for companionship and care for older people, providers of information in government administration or screening and detection of diseases thanks to telemedicine are some of the applications foreseen.

Spoken Language Understanding (SLU) is an emerging field between speech recognition and natural language understanding, where the objective is to extract structured information from the speech signal. SLU systems are one of the main components of a VA, and they are traditionally designed as a pipeline with an ASR module to convert the speech into text, followed by an NLU module, where the inferred text is analysed to extract structured information, typically as a set of domain, intent and slots or entities. However, these systems are optimised independently and under different criteria: the ASR module is trained to reduce the word error rate (WER) criterion, where each word weighs the same, and the NLU module is usually trained with clean text, without taking into account the errors at the ASR output, especially in noisy conditions.

End-to-end (E2E) learning has been used in several areas with success (ASR, machine translation, speech synthesis, etc.) and for years now E2E SLU architectures have been a hot topic in research (Serdyuk et al., 2018). In E2E SLU the structured information is extracted directly from the speech signal, so they are more efficient and compact systems, where optimisation is performed directly on the target task, and where the cascading errors of traditional pipelines are avoided. These systems will allow to develop more efficient solutions that, for example, enable edge computing (Edge AI), avoiding sending voice recordings to cloud computing networks or private data centers. Nevertheless, E2E SLU modules have limitations related to the fact that data needed to train must be specifically collected (speech data) and the datasets need to be carefully designed for the task. Most European languages do not have the necessary resources to train such systems and, currently, Big Tech companies are the ones that mainly have access to the necessary resources: because of their developments in smart speakers or conversational assistants and because the datasets must be specifically designed for these tasks, which makes them expensive.

As it is been set out in the The Strategic Research and Implementation Agenda (SRIA) (ELE-

Consortium, 2022), handling data scarcity when working with speech technologies is one of the major challenges for European languages (as well as the privacy problem). This project aims to make it possible to create such systems in any of the European languages, especially in low-resource scenario, establishing criteria for the design of datasets, their acquisition, their validation and presenting the lessons learned in the process. Furthermore, by making this data available in a liberalised form, it is possible to the development of the virtual assistants of the future from the economic interests of the Big Tech companies, developing multilingual virtual assistants in fields that are not profitable for them, but can be for the European population.

This project has been carried out through a collaboration between Balidea Consulting & Programming S.L.[1] and the Multimedia Technologies Group of AtlanTTic[2], from the University of Vigo (UVIGO). Balidea, founded in 2002 and with headquarters in Santiago de Compostela (Spain), has more than 20 years of experience in IT Consulting Services and Software Engineering in different sectors, but its main projects are related to e-Health and Public Health Area. Nowadays, Balidea is working on research projects focused on Natural Language Understanding, Natural Language Generation and Spoken Language Understanding, developing technologies that facilitate communication in Galician and Spanish through virtual assistants. On the other hand, the GTM research team of the University of Vigo has extensive and proven experience (more than fifteen years) in the research lines of text-to-speech conversion (TTS) and automatic speech recognition in Galician and Spanish, and in the line of analysis and facial recognition (started more than 25 years ago). In these fields, it has participated in a number of projects funded at European, national and regional level, in addition to various contracts with companies. As a result, this team has its own algorithms and software tools, which can be adapted to a new experimental framework in a short time.

In short, this project aims to contribute to the SRIA by presenting a guidance for designing and collecting datasets for E2E SLU systems, taking advantage from Balidea´s experience in e-Health virtual assistants. The objectives are to:

1. Guide on how to design SLU datasets in low-resource scenarios, establishing the required characteristics and proposing design quality measurements.
2. Guide on how to design data collections campaigns, focusing on target users.
3. Present lessons learned with the campaigns.
4. Present methodologies to validate collected datasets.
5. Present quality measurements over collected datasets.

In this way, we will contribute to the success of the strategic agenda by establishing guidelines on how to approach an E2E SLU project in a low-resource scenario and, responding to the agenda's concern about how technological progress benefits in an equitable and fair way, we will share previous experiences with the community to achieve digital equality between European languages, paying special attention to those coexisting with a major language. To do this, section 2 presents the task and the recording tool, section 3 presents the dataset design and the complexity measurements, section 4 presents the campaign design, section 5 and section 6 present the campaign evolution and the validation methods and, finally, the section 7 and the section 8 present the lessons learned and the conclusions of the project.

[1] https://balidea.com/en
[2] http://gtm.uvigo.es/en/

2 Motivation: Creating and obtaining a dataset for SLU from scratch

In order to train E2E SLU systems, it is necessary to have specifically designed speech datasets and domains. Traditional task-oriented datasets such as automatic speech recognition, speaker identification, or automatic subtitling are not useful for the SLU task (or at least not sufficient to train these systems). In E2E SLU, structured information is extracted directly from the speech signal, usually in the form of domain, intent and slots, so it is necessary that the sentences in the corpus follow this information structure, i.e., the sentences in the dataset need to be requests, commands or questions of some kind, and that the structured information is correctly labelled as metadata. This, together with the fact that, compared to text, it is not possible to generate unlimited speech data, and that it is therefore more complex to obtain representative speech datasets, makes it very costly to create and obtain datasets for E2E SLU. The main commercial applications of virtual assistants have meant that Big Tech and its labs are currently leading research in this field, and therefore it is mainly these companies that have the resources to develop this technology and are also the ones that decide on the applications of the technology.

It is necessary not only to have the speech data resources, but also to have the knowledge and the necessary guidelines to design and create this type of datasets, allowing access to technology to languages with low-resources, and decoupling the development of virtual assistants from the economic interests of large companies. In this way, it will be possible to develop technology for fields that are not economically profitable, but are important on a social level, and in fields where access requirements prevent many Big Tech solutions (as they are not auditable or have total governance over the data), such as healthcare, education or public administration.

2.1 Task definition

To achieve the goals of the project, we have taken advantage of the current work and research framework of Balidea in end-to-end SLU systems. Balidea, in collaboration with the University of Vigo, is involved in projects of conversational voice assistants for e-health applications, with the target users being older people who speak Galician and Spanish. As part of these projects, Balidea has also developed a tool for the collection of speech recordings, which will be the basis of this project and which will be reused in the collection campaigns. Within this framework, and with the ultimate goal of creating and collecting an SLU dataset from scratch, the following tasks have been defined within the project:

- **Task 1**: Perform a study on the minimum design features of a SLU dataset for low-resource scenarios based on the experience with bilingual voice-based VA.
- **Task 2**: Propose quality measures, regardless of the language of application, to determine the complexity of the designed dataset, in order to be able to establish minimums in the design and collection of data.
- **Task 3**: Design the data collection campaigns, which will be oriented towards weak language data, and trying to record the main language varieties.
- **Task 4**: Present the lessons learned and the results of the data collection campaigns.
- **Task 5**: Present the methods of validation of the collected data, as well as series of quality measurements over the final dataset.

2.2 The idea of the project

Many European languages are currently suffering from a scarcity of resources, so projects that lay the foundations of how to involve collective participation are very important, and from which all other languages in the same situation can benefit. Although this is a multi-objective project with several defined tasks, its success depends mainly on the success of the voice data collection. To properly understand the importance of data collection and the success of the campaigns, it is necessary to know the situation of the language under study, Galician.

Naturally, in line with the direction set by the SRIA, all the resources obtained in this project will be available for other research groups or companies to continue researching or developing speech technology in Galician. Section 2.3 specifies in detail the conditions under which participants have made their recordings available.

Galician context

In order to properly understand the public approach of the project, it is necessary to know the context of Galician. Galician, which belongs to the Romance language family, is a co-official language, together with Spanish, in the autonomous region of Galicia, located in northwestern Spain, and has approximately 1.9 million speakers (I.G.E., sep 2019a). Its use also extends to neighbouring provinces such as Asturias, León, Zamora or three municipalities in Extremadura, and due to emigration, large concentrations of Galician-speakers can also be found in other regions of Spain (Madrid, Barcelona, the Basque Country and Canary Islands), Europe (Portugal, France, Switzerland, Germany, United Kingdom, Netherlands) and America (Argentina, Uruguay, Brazil, Venezuela, Cuba, Mexico and the United States) (O'Rourke, 2014). Moreover, Galician is a language that have linguistic variations depending on the geographical area, and which coexists in a situation of bilingualism and code-switching.

Despite its rich cultural tradition and the fact that it is the official language in public institutions, the digital presence of Galician is scarce (Sánchez et al., 2022). As stated in the ELE report on Galician (Sánchez and Mateo, 2022), it belongs to the group of languages with fragmentary support, but it is a borderline case. The lack of resources clearly affects the development of language technologies (LTs) such as automatic speech recognition, natural language processing, machine translation, text analytics or dialogue systems. Efforts are currently being made to reverse this situation, and as part of the Spanish Plan for the Advancement of LT[3], the Galician government has presented the Nós project (*Proxecto Nós*) (de Dios-Flores et al., 2022), 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. In order to make efficient use of available resources and to be as effective as possible, it is important that such projects (public or private) are coordinated with other ongoing projects. This aspect will be addressed as part of the lessons learned in the section 7.

For the specific case of end-to-end spoken language understanding, there is no existing dataset of speech recordings for Galician, so the dataset collected will be the first public dataset of these characteristics. If we analyse the available speech resources (although they are not suitable for E2E SLU, they can be used with traditional ASR + NLU architectures), the resources available in Galician are also scarce. The two main datasets available are the Common Voice dataset (Ardila et al., 2019), which in its version 12[4] has a total of 17 hours of validated speech, and the openSLR dataset (Kjartansson et al., 2020), which has approximately 10 hours of audio, so that in Galician there are a total of 27 hours of quality speech available.

[3] https://plantl.mineco.gob.es/Paginas/index.aspx

[4] This is the latest version available at the date of creation of this deliverable.

FalAI

Since the success of such a project depends on the altruistic participation of the population, it is necessary to give the project a very concrete and attractive image. In order to achieve this, the first necessary change was to give the project a "commercial" name and, in addition, it is also necessary to narrow down the information provided to the population and to the media, in order to make the project more understandable, more accessible, and more attractive. The name chosen was **falAI**, which is a composition between the Galician verb for speak, "falar", and the acronym for artificial intelligence, AI. Moreover, in some areas of Galicia it is how the second-person plural of the imperative of the verb to speak is created: "falai" instead of "falade". Throughout the deliverable, reference will be made to falAI as the project to collect voices that the population perceived and campaigns used.

FalAI was advertised as a research project of the University of Vigo (in the section 7 we will emphasize the importance of the participation of public organizations) and the company Balidea for the collection of voices to train artificial intelligence models for conversational assistants in Galician. The Galician population is perfectly aware of the lack of technological tools in Galician (there are no conversational assistants with a Galician option, there are no spell checkers in Galician, many cell phones do not have the possibility of configuring Galician as a language, etc.), so there was no need to explain why this type of action is necessary. To sell the project it was only necessary to establish a clear image, clear objectives and a suitable tool.

2.3 FalAI voice collection tool

In addition to an attractive project image and clear and concise information and objectives, in order to get people, regardless of their age or technical background, to participate in falAI, it is essential to have a multi-platform, simple and well-functioning speech recording tool. For the voice collection campaigns, a tool previously designed by Balidea for this purpose has been used. The tool was simply updated to include the appearance and images of the project, the project information, and some additional functionalities that were necessary due to the campaign strategies. It is a tool accessible through a URL[5], designed to be displayed correctly on PCs but especially on mobile devices, as a greater participation from this type of devices was expected. All the necessary interactions with the tool are carried out through buttons or drop-downs, seeking greater simplicity in its use. In addition, special emphasis was placed on creating a tool that would appear intuitive for a non-technological profile, using an appropriate font size, avoiding technical terms, and including the necessary instructions.

Figure 1(a) shows the initial screen of the falAI tool. It shows the falAI logo, a brief description of the project[6], information about the possibility of winning a prize (this part will be discussed in section 4.2), and a button to start the recording process.

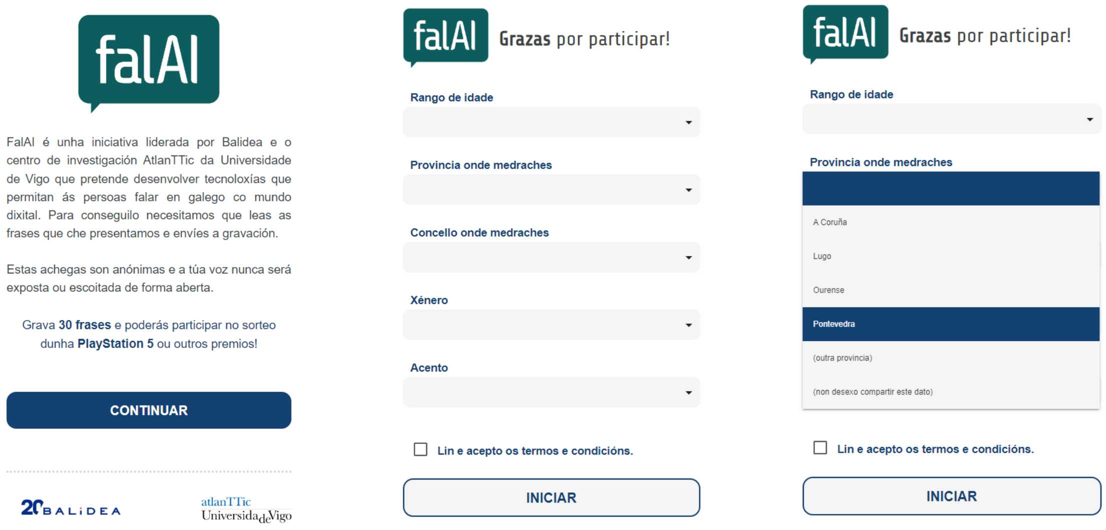
The next screen of falAI is shown in the Figure 1(b). In this screen the user optionally indicates their age range, where they grew up (where they learned the language), gender and accent. In order to start the recording process, it will be necessary to affirm that the terms and conditions of the assignment have been read and accepted. It will not be possible to tick the checkbox without having opened the document. The document on the cession of voice rights[7] has been drawn up on the basis of the GDPR and with the guidance of legal experts.

[5] **FalAI recording tool**, accessible via: https://falai.balidea.com

[6] The translation would be: falAI is an initiative led by Balidea and the AtlanTTic research centre of the University of Vigo that seeks to develop technologies that allow people to speak Galician with the digital world. To achieve this, we need you to read the sentences we present and send the recording. These contributions are anonymous and your voice will never be exposed or heard freely. Record 30 sentences and you will be entered to win a PlayStation 5 or other prizes!

[7] Accessible via: https://falai.balidea.com/privacy-gl/privacy_policy_gl.pdf

It sets out what Balidea is and the terms of cession (anonymous, worldwide territorial scope, for an indefinite period of time and without limitations for the data to be used in other research or technology development projects). Furthermore, the user must declare that he/she is of legal age and has the legal capacity to grant the rights granted in the document. Once you have completed all the data, you can start the recording process.



falAI

FalAI é unha iniciativa liderada por Balidea e o centro de investigación AtlanTTic da Universidade de Vigo que pretende desenvolver tecnoloxías que permitan ás persoas falar en galego co mundo dixital. Para conseguilo necesitamos que leas as frases que che presentamos e envíes a gravación.

Estas achegas son anónimas e a túa voz nunca será exposta ou escoitada de forma aberta.

Grava **30 frases** e poderás participar no sorteo dunha **PlayStation 5** ou outros premios!

CONTINUAR

(a) Initial screen

falAI **Grazas** por participar!

Rango de idade

Provincia onde medraches

Concello onde medraches

Xénero

Acento

☐ Lin e acepto os termos e condicións.

INICIAR

(b) Parameter list

falAI **Grazas** por participar!

Rango de idade

Provincia onde medraches

A Coruña
Lugo
Ourense
Pontevedra
(outra provincia)
(non desexo compartir este dato)

☐ Lin e acepto os termos e condicións.

INICIAR

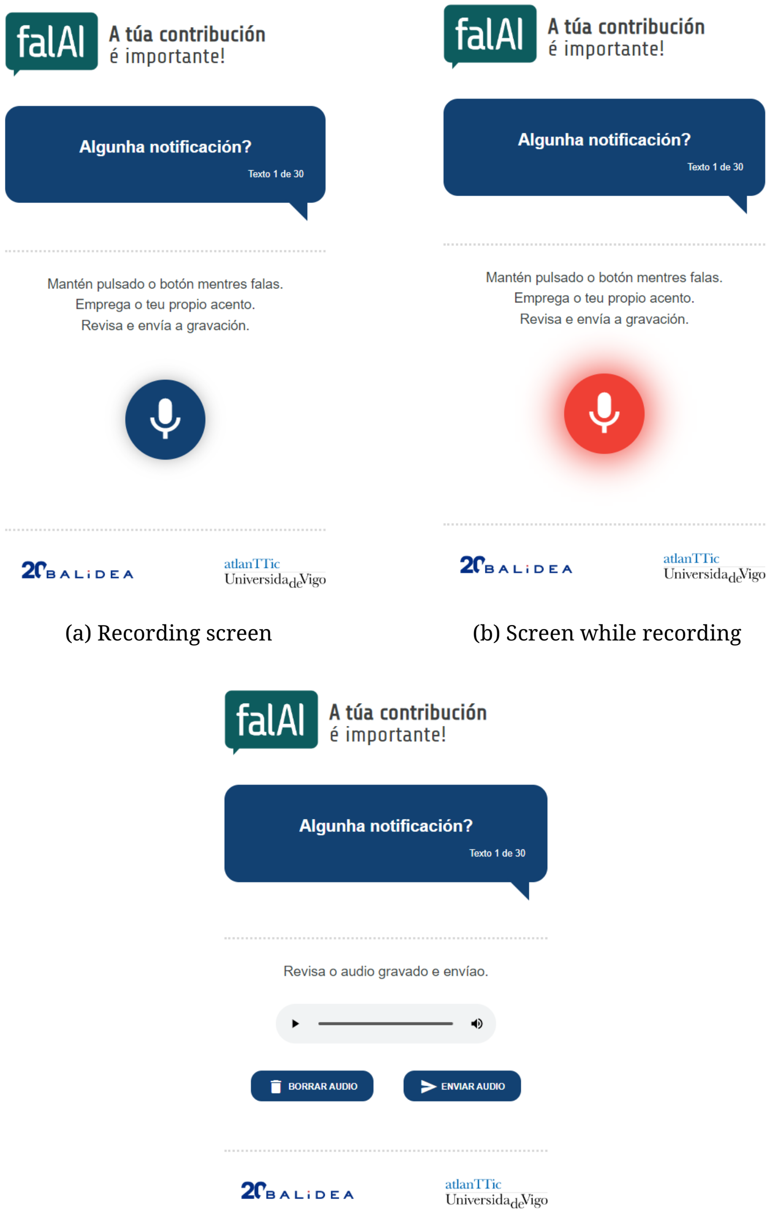
(c) Drop-down example

Figure 1: Initial screen and information entry screen of falAI.

The sentences recording screen is shown in Figure 2. It shows the sentence to be read on a blue background and in a larger font size, the instructions, which indicate to hold down while speaking, to use one's own accent and style of speech (something that will be strongly emphasised during the campaigns: participants must use their own way of speaking and their own accent, and may even change words in the sentence), to review the recording before sending it, and the button to press while reading (microphone symbol). It has been decided to segment the number of recordings by the user in series of 30 sentences (we have tried to reach a compromise between asking the participants for as many sentences as possible and that the total time needed to participate should not exceed 5 minutes). This will be explained in detail in section 4.2). The number of sentences within the series is always displayed on the screen, in order to motivate the completion of the 30 sentences.

While recording, an animation is prepared that changes the colour of the microphone to red, in order to make it more intuitive and clearer (Figure 2(b)). Once the recording is finished, it will be possible to review it, and in case it is not correct, delete it and re-record it (Figure 2(c)). In order to move forward in the process, the recording needs to be accepted by a specifically developed automatic validator (explained in detail in the security measures section)

Likewise, the algorithm for selecting sentences within the corpus works as follows: the tool searches for sentences with fewer recordings, and at the start of the recording process it randomly selects the 30 sentences that will be shown to the user. In this way we ensure that the contributions per sentence are balanced.


falAI
A túa contribución é importante!
Algunha notificación?
Texto 1 de 30
BALIDEA
atlanTTic Universida de Vigo

Mantén pulsado o botón mentres falas.
Emprega o teu propio acento.
Revisa e envía a gravación.

(a) Recording screen

falAI
A túa contribución é importante!
Algunha notificación?
Texto 1 de 30
BALIDEA
atlanTTic Universida de Vigo

Mantén pulsado o botón mentres falas.
Emprega o teu propio acento.
Revisa e envía a gravación.

(b) Screen while recording

falAI
A túa contribución é importante!
Algunha notificación?
Texto 1 de 30
BORRAR AUDIO
ENVIAR AUDIO
BALIDEA
atlanTTic Universida de Vigo

Revisa o audio gravado e envíao.

(c) Review after recording

Figure 2: Recording screen of falAI.

Data schema

A MongoDB database has been used to store the information. Although there are more tables, the structure of the two main ones is shown below.

Sentences table Table with information on the sentences of the corpus and the list of associated recordings. Each audio has an associated user, which is created at the start of the recording process:

```
{
    ”_id” : ObjectId(”6405ef5bb363c53c8cb1dae9”),
    ”sentenceid” : ”afe2517754af4245abcb20c13f703e53”,
    ”audios” : [
        {
            ”userid” : ”eeaf3010448340a5b5b3a94ac928c145”,
            ”duration” : 4.3653514739229,
            ”fileid” : ObjectId(”6406047a9ed3f8fb721565d7”),
            ”date” : ISODate(”2023/03/06T15:19:22.170Z”),
            ”reviewed” : true,
            ”status” : ”validated”,
            ”transcription” : ”quero escoitar algunha pansoriña en galeou”,
            ”cer_ratio” : 0.0952380952380952,
            ”hash” : ”ad429d5ef8fc8b13099bd8042d29a77855970dd3f6”
        },
        {
            ”userid” : ”bbd5b216cd00428f84b58767ae437486”,
            ”duration” : 3.15791383219955,
            ”fileid” : ObjectId(”6416181e1a9f445aa3f78eb5”),
            ”date” : ISODate(”2023/03/18T19:59:26.864Z”),
            ”reviewed” : false,
            ”status” : null,
            ”transcription” : ”quero escoitar algunha panxoliña en galego”,
            ”cer_ratio” : 0.0238095238095238,
            ”hash” : ”4359fe0df8850a83e309c1019d0e7f019b243e8355”
        }
    ],
    ”date” : ISODate(”2023/03/06T13:49:15.461Z”),
    ”domain” : ”house commands”,
    ”intent” : ”music”,
    ”model” : ”tese”,
    ”sentence” : ”Quero escoitar algunha panxoliña en galego”,
    ”slots” : [
        ”action” : ”play”,
        ”type” : ”panxoliña”
    ]
}
```

Listing 1: Sentence schema example.

Users table The user table with the information provided by the user before starting the recording process:

```
{
    ”_id” : ObjectId(”62fdf1713b12773e2ac0a2de”),
    ”userid” : ”4fcbdc4d498a4629a73f84488d3f1f04”,
    ”accent” : ”central”,
    ”accept_conditions” : ”si”,
    ”age_range” : ”range50_59”,
    ”gender” : ”masculino”,
    ”location” : ”32004”,
    ”province” : ”32”
}
```

Listing 2: Users schema example.

Security measures

As the project was expected to be widely disseminated throughout the population, it has been decided to implement a series of security measures to try to reduce the chances of attacks on our servers, denial of service (DoS) or malicious access to the benefits of completing the 30-sentence series (explained in section 4.2).

Balidea's servers, where falAI is deployed, are already protected against DoS attack attempts and against repetitive and suspicious requests from the same IP. The operation of falAI at the client-server level also reduces the chances of attacks. When a client starts the recording process, it is given an *execution_id*, which will only be valid for a series of 30 sentences. In addition, the endpoints where falAI requests are made are secured through a token, which is linked to the *execution_id*. This token expires after 30 sentences.

Finally, a validator was implemented, which consists of an ASR that calculates the CER (Character Error Rate) between the reference sentence and the inferred one. In case the CER does not exceed a minimum threshold, the message "Sorry, I didn't understand you correctly" will be displayed and the user will be asked to repeat the recording. It has been decided to use the CER and not the WER (Word Error Rate) because it is a more robust measure for this type of validation, where you simply want to check that you have recorded a sentence similar to the reference sentence, and because in such short sentences the WER may not be as robust as the CER. Checking the CER of the transcript for each recording sent is an important security measure, as it reduces the possibilities of sending ghost recordings, the possibilities of sending sentences other than the reference sentence, and the possibility of creating bots that automate the process in order to access the benefits of completing sets of 30 sentences (section 4.1 and section 4.2).

3 Dataset design

There are many aspects to take into account when designing the SLU dataset. The objective is to design a dataset that has the minimum features to be useful at a research level, but also at a practical level in real scenarios. In addition, the particularities of the language must be taken into account, as well as the possible linguistic variations that also want to be collected. In this section, the characteristics of the main SLU datasets are analysed, complexity measures are proposed, and the characteristics of the dataset designed for falAI are presented and compared.

3.1 Benchmark SLU datasets

The first corpus in the literature that contained audio and annotated semantic information was the Air Travel Information System (ATIS) (Hemphill et al., 1990), introduced in the 1990s. It was not until the first end-to-end approaches for SLU were introduced (Serdyuk et al., 2018) and (Haghani et al., 2018) that the first E2E SLU corpora in English were created: The SNIPS benchmark (Coucke et al., 2018) and the Fluent Speech Commands (FSC) (Lugosch et al., 2019) corpus, which became the main corpus on which to compare state-of-the-art results. However, the results obtained in these corpora were not representative of the actual performance of the technology, as they were small corpora with low semantic complexity (McKenna et al., 2020). It is for this reason that efforts to create and share E2E SLU corpora continued. The next large corpus introduced was the Spoken Language Understanding Resource Package (SLUPR) (Bastianelli et al., 2020), which contains 6 times more sentences than SNIPS, 2.5 times more audio than FSC, with more domains and greater lexical richness. The largest SLU corpus in the literature to date, the STOP datasets, was introduced in early 2023 by Meta (Tomasello et al., 2023). This dataset was a quantitative step forward in terms of the number of audios (three times more audios than SLURP) and the number of speakers (five times more than SLURP), but also includes compositional queries with nested intents, which no previous publicly available SLU datasets included (introduced in TOP dataset for NLU (Gupta et al., 2018)). The table 1 shows a comparison between the text characteristics of the main SLU datasets and **falAI dataset**.

	FSC	SNIPS	SLURP	STOP	**falAI**
Phrases	248	2,912	17,181	125k	**3,500**
Domains	1	1	18	8	**14**
Intents	31	7	46	80	**62**
Slots	-	53	55	82	**64**
Vocabulary size	96	2,182	6,467	15,056	**2,957**

Table 1: Text corpora SLU dataset comparison.

3.2 Dataset complexity

There is currently no benchmark measure for the complexity of an NLU or SLU dataset, although there are widely used indicators that can be quantified using descriptive statistics such as mean, median, standard deviation, and variance. Additionally, visualization techniques such as histograms and bar charts can be used to represent these measures visually. Some of the indicators are:

1. **Dataset size**: Quantified in terms of the number of examples (e.g. sentences) in the dataset.
2. **Level of natural language ambiguity**: Quantified in terms of the number of ambiguous sentences in the dataset.
3. **Variety of topics and domains**: Quantified in terms of the number of topics and domains covered in the dataset.
4. **Diversity of slots**: Quantified in terms of the number of different slots in the dataset.
5. **Label ambiguity**: Quantified in terms of the number of ambiguous or poorly defined slots in the dataset.

It is well known that near-perfect results obtained with datasets such as FSC or SNIPS are not representative of real scenarios, because they have limited corpora in terms of lexical or semantic richness, number of vocalisations, domain coverage and semantic contexts (Bastianelli et al., 2020). Studies have been carried out to try to establish measures of semantic complexity (McKenna et al., 2020), which could be used in the design of the dataset. The aim of this section is to compile the main metrics of the corpora, which allow us to evaluate its complexity, and to propose new metrics based on our experience.

3.2.1 Lexical analysis: n-Gram Entropy

In addition to the vocabulary size (number of unique words in the corpus of sentences) and the number of unique sentences, a good measure of lexical complexity is the n-gram entropy. The n-gram entropy measures the randomness of the dataset sentences over its constituent n-grams, $\mathcal{N}$. It is calculated by the equation:

$$H = -\sum_{x \in \mathcal{N}^*} p(x) log_2 p(x)$$

where $\mathcal{N}^*$ is the set of unique n-grams in the dataset and $p(x)$ is the probability of n-gram x occurring in $\mathcal{N}$. Larger values of n-gram entropy represent higher randomness and variety in the utterance patterns, indicating larger lexical complexity. Table 2 shows a comparison between the entropy for unigrams, bigrams, trigrams and average entropy between what has been the main benchmark datasets in the literature to the date, the Fluent Speech Commands dataset and the SNIPS dataset, and the falAI dataset.

Entropy	FSC	SNIPS	**falAI**
unigram	5.5	6.2	**9.26**
bigram	7.2	9.1	**12.49**
trigram	7.9	10.9	**13.38**
average	6.9	8.7	**11.71**

Table 2: Comparison of entropies between the main SLU datasets and falAI dataset.

3.2.2 Syntactic analysis: lexical sophistication (LS2)

Lexical sophistication (LS) is a measure of the complexity and diversity of vocabulary used in a text. It is often used as a metric for evaluating the quality and sophistication of written or spoken language. There are several ways to measure LS, but one common approach is using the Lexical Sophistication 2 (LS2) measure (Laufer, 1994).

Lexical sophistication has been compared in great detail in SLURP (Bastianelli et al., 2020) for the benchmark datasets, and there are currently no tools available for the calculation in Galician. Moreover, comparing the lexical sophistication between different languages can be a challenging task. It is possible to compare the size of vocabularies across languages, but simply comparing the number of words in a language may not provide an accurate representation of the lexical sophistication of that language. The complexity and richness of a language depend on various factors such as the grammatical structure, the syntactic complexity, the semantic range, and the cultural context of the language. Overall, while it is possible to compare certain aspects of the lexical sophistication between languages, it has been decided not to compare them as it is a complex and nuanced task that requires careful analysis and consideration of various factors.

3.2.3 Semantic analysis: semantic textual similarity (STS)

Semantic Textual Similarity (STS) measures the degree to which two sentences are semantically equivalent to each other (Cer et al., 2017). This task has typically been used in applications such as machine translation, summarization, question answering or semantic search. In this section we propose to calculate STS as a measure of the semantic complexity of the designed dataset.

To calculate the STS, first the vector representations (embeddings) of the sentences in the corpus are obtained. For this purpose, the Language-agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2020), based on BERT (Devlin et al., 2018) and trained for more than 100 languages (including Galician, Spanish and English), has been used. After this, an L2 normalisation (Feng et al., 2020) is performed to avoid that parameters of the vectors with different ranks and a high variance affect more than those that are not normalised, and then the similarity is calculated as the product between the two tensors.

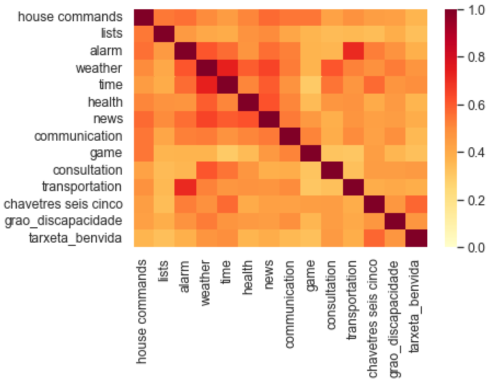
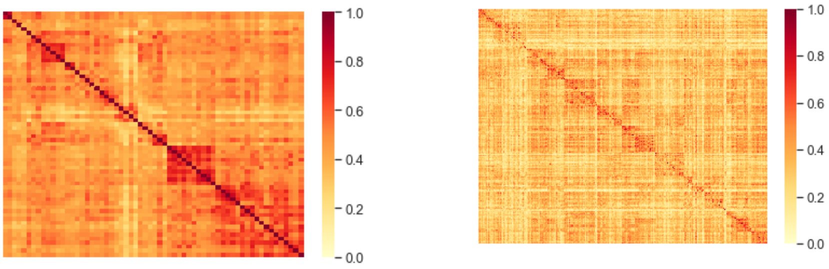


Figure 3: STS between domains in falAI calculated with LaBSE.



(a) STS between intents

(b) STS between sentences

Figure 4: STS between intents and sentences in falAI calculated with LaBSE.

Figure 3 shows a heatmap of the STS between the falAI domains, with 1 being the highest similarity and 0 the lowest similarity. It can be seen that there is some similarity between some of the defined domains mainly related to time, weather or alarms. If we analyse the results in terms of STS between intents (Figure 4(a)), we can see that there are more intents with greater similarity, which can complicate the correct classification (higher complexity). As expected, the similarity is lower if we analyse at the sentence level (Figure 4(b)), obtaining 0.12 as average correlation.

3.3 FalAI dataset features

The text corpus, designed in the first month of the project, had an initial version of 1,500 sentences and over 8 domains. Due to the success of the campaigns, once the project started, it was decided to increase the corpus to take advantage of the opportunity. This section presents the characteristics of the final dataset.

Within the domains defined, the one with the most sentences is health. This is because Balidea expects to be able to use this data in real conversational assistant projects, user case announced in the project proposal. Among the defined domains are health, typical conversational assistant domains (house commands, music, events), domains that involve integrations with other systems (appointments, transportation) or information domains in administrative processes. In addition, most of the defined slots are highly complex, as they refer to Galician municipalities, personal names or specific locations.

The corpus has been created in collaboration with linguists in order to try to collect all the variants of the language, to correct and revise possible errors introduced and to establish criteria for its creation. Some of the criteria established are:

- All the words in the corpus must be words accepted by the *Real Academia da Lingua Galega*, the institution in charge of establishing the rules for the correct use of the language. Although there are many variants of the language and widely used terms not covered by the institution, this criterion has been established because of the complexity that would be involved in establishing criteria outside the norm and because of the time constraints of the project.
- An effort has been made to balance gender and place references within the corpus, trying to avoid bias in the creation of the corpus. We have also tried to involve as many people as possible in its creation, increasing diversity.
- In Galician there are situations in which two types of writing are valid, but only one of them can be pronounced, i.e. there are situations in which it must be pronounced differently from the way it is written[8]. Currently, the tendency is to write the form that cannot be pronounced, and this is how the corpus began to be designed. However, when analysing the first recordings, it was found that most of the participants did not pronounce correctly, as they read literally what was written. For this reason, and after consultation with the team of linguists, it was decided to introduce the two written versions into the corpus. In this way we do not limit the user's pronunciation, we ensure that there are sentences where it is pronounced according to the norm, and we create a corpus that can be very interesting at a socio-linguistic level in the future.

3.3.1 Minimum design features

Based on Balidea's experience with conversational assistants in bilingual contexts, and on the basis of the first tests of end-to-end SLU systems, it is necessary to have datasets with a higher complexity than those that existed until recently in the literature (ATIS (Hemphill et al., 1990), SNIPS (Coucke et al., 2018) or FSC (Lugosch et al., 2019): their almost perfect results already indicated that they would not work well in real scenarios). Datasets such as SLURP (Bastianelli et al., 2020) or STOP (Tomasello et al., 2023) are much more complete and could be extrapolated to production scenarios. As part of this project, an attempt has been made to design a dataset that can operate correctly in a production scenario on the application domains considered. Based on our experience, a limited conversational assistant

[8] For example, the following sentences are correct in written form, but only the second one is correct when read: *Activar o radiador / Activa-lo radiador.*

usually has a number of intents between 15 and 40, and they are usually single or limited domains. If the corpus is correctly designed, it is possible to create systems with adequate performance with a minimum of 50-100 sentences for each intent and each language (varying according to the complexity of the classification, the number of slots or the number of domains).

Although it cannot be tested within the framework of the project, the collected dataset will be used to study the viability of this type of dataset in real projects, establishing minimum requirements in the design of the assistants. The impact of using data augmentation and synthetic data in this type of solution will also be studied, which may modify the minimum requirements. In practice, it would be common to start with a small amount of training data and then iteratively add more data and refine the model based on its performance. This approach allows to gradually improve the accuracy of the model over time.

4 Campaigns design

As discussed in the objectives section, the success of the project depends on the success of the data collection, so the design of the campaigns is of vital importance. Although the lessons learned and strategies presented can be applied in other scenarios, it should be noted that due to the time constraints of the ELE project itself, the duration of the campaigns had to be adjusted to the deadlines of the deliverables. The design of the campaigns has been carried out in collaboration with Balidea's communication team, the communication team of the University of Vigo, and an advertising company contracted by Balidea. Some key aspects for the data collection campaigns design:

- KPI campaign objectives: Table 3 shows the KPI-level objectives defined for the campaign, which have been estimated on the basis of the average duration of the first recordings and the proportion of the population in each province.
- Informed consent: As explained in section 2.3, in order to participate in falAI it will be necessary to accept a data release document, which informs how the data will be used, who will have access to it, and how the data will be protected. In addition, an effort will be made to explain that the data obtained in the campaign will be openly available for research and technology development in Galician, and that voice data will be treated with utmost respect and ethics. This means ensuring that data is stored securely and protected from possible leaks or misuse.
- Participant selection: A segmentation of the target groups of interest was carried out in order to be able to use different strategies and check their effect. The campaign aims for mass participation, but also for some of the strategies to have a special impact on older people.
- Speaker diversity: The campaign aims to get speakers from all parts of Galicia, with representation of linguistic varieties, accents, with a similar participation of both sexes and with the presence of all possible age ranges.

Real time data dashboard

In addition to the voice recording tool, it has been decided to implement a dashboard with real-time campaign data[10]. The aim is to be able to monitor the performance of the campaign

[9] Galicia has four provinces: A Coruña, Lugo, Pontevedra and Ourense.

[10] **Dashboard created in Metabase and accessible via: https://falai.balidea.com/datos-tempo-real/public/dashboard/c2d965e3-287d-4187-a95e-482fbe82578f/**

	KPI estimation
Total samples [hrs]	100
Samples from people older than 50 years [hrs]	30
Samples from each province[9] [hrs]	10
Participants	6,000

Table 3: KPI objectives estimation for campaigns.

internally, evaluate the results and analyse trends, but also to create this dashboard as a campaign tool for the participants. We wanted to provide participants a tool that would allow them to check the evolution of the campaign with respect to the objectives set, the municipalities with the highest participation, or the accents or age ranges with the highest representation, which also serves to motivate the participation of the population, either to achieve the set objective or as a competition between municipalities, accents or age groups. Figure 5 shows progress bars on the defined objectives[11]:

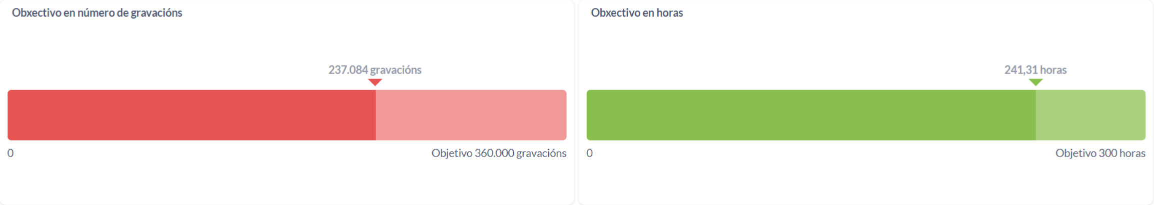


Figure 5: Objectives in terms of number of recordings and number of hours.

Figure 6 shows a map with the distribution of participants among the 313 municipalities in Galicia and a ranking of the 12 municipalities with the highest number of participants:

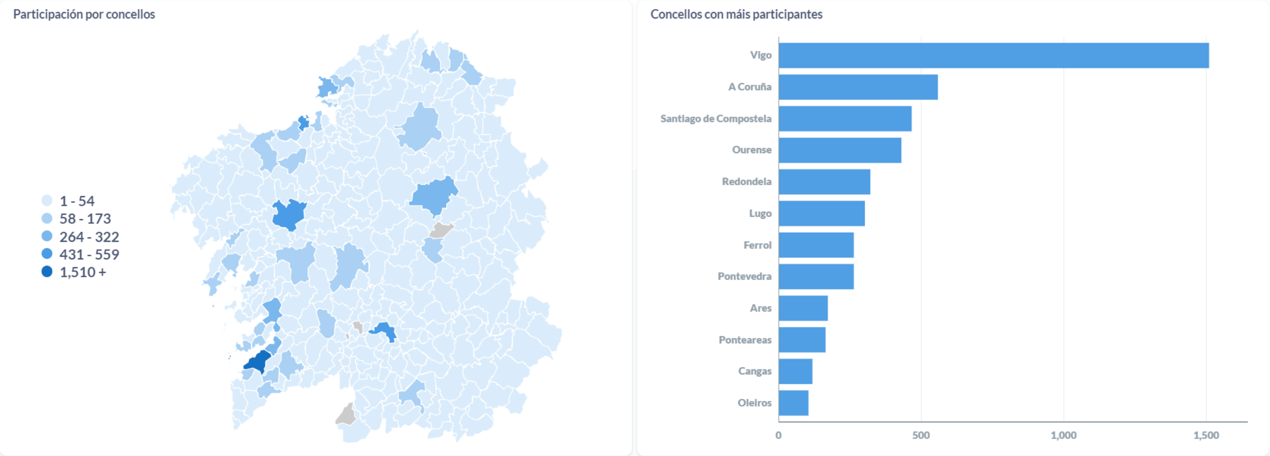


Figure 6: Galicia map with distribution of participants and municipalities with the highest number of participants.

[11] At the time these images were taken, the initial objectives had already been modified due to the success of the campaign.

Figure 7 shows a graph with the number of participants per province and the distribution according to shared gender, while figure 8 shows the distribution of participants according to accent, and a motivational graph with the number of hours achieved.

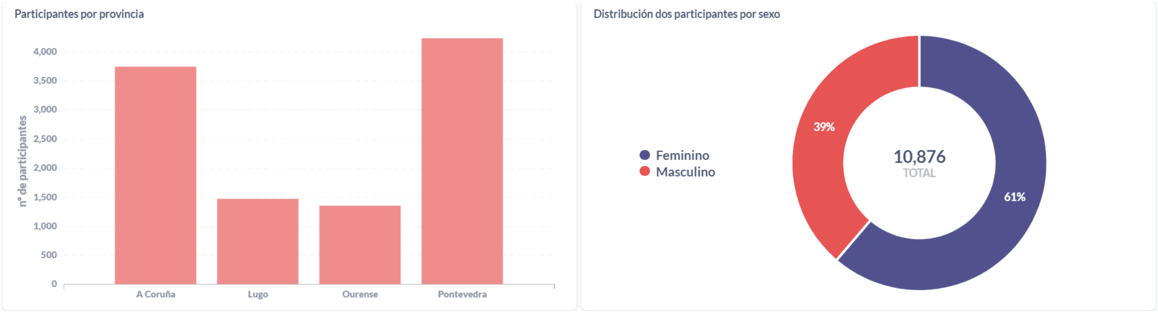


Figure 7: Number of participants per province and the distribution according to gender.

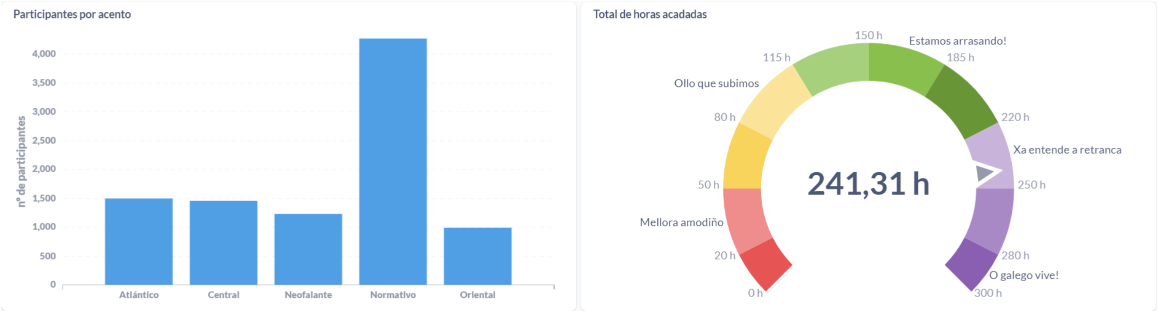


Figure 8: Distribution of participants according to accent and a motivational graph with the number of hours achieved.

4.1 Media campaign

The media campaign is the most far-reaching campaign and the one that will in parallel with the other campaigns. The main objective is to promote falAI in all the channels available to us, encouraging public participation and creating a clear brand image. The campaign will seek to make as much noise as possible with a presentation with public personalities and seeking institutional support, interviews in newspapers, radio and television, advertising on social networks, creating falAI profiles on social networks and collaborating with local influencers and celebrities. The profiles, created on Instagram[12] and Facebook[13], are intended to be the platform where falAI's content can be displayed without having to link the University of Vigo or Balidea directly, which gives greater flexibility for its creation. It will be from these profiles that collaborations and support from influencers and celebrities will be requested.

To complement all the promotion, digital content has been created (which will also be used in the rest of the campaigns) where the project is explained and promote, and will serve as support for all the information provided. In addition, a slogan has been created around

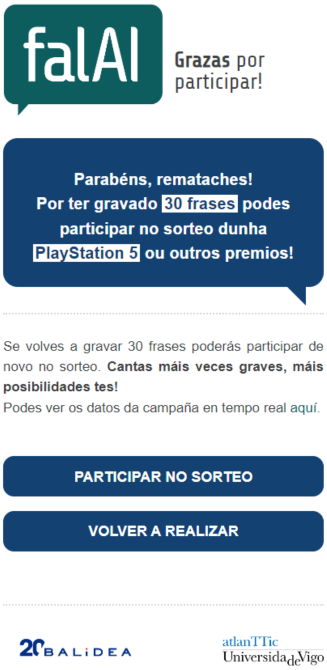
[12] https://www.instagram.com/retofalai/
[13] https://www.facebook.com/people/Reto-Falai/100090243069834/

which most of the promotion of falAI will be based. This slogan seeks to motivate the entire population that wants Galician to be a language with technological representation: you are 30 sentences away[14] of achieving that the machines understand Galician. The idea is to set up something like a challenge to motivate people.

The fact that there are 30 sentences has been discussed and planned. We wanted people to participate and record more than one sentence, but without asking for too many and making the process tedious. The final number was chosen based on the average time it takes to complete a series: the average length of a recording is 3.5 seconds, and leaving a margin between recordings, we wanted the total time to complete to be between 3 and 5 minutes, no more, so we came up with this number of 30 recordings, in which we ask for significant collaboration from people but without causing tedium or boredom. It should be noted that longer and shorter blocks were tested, and the 30-sentence block was the one that gave the best results without causing significant irritation.

4.2 Young people campaign

Campaign targeted at young people over the age of 18 (you must be of legal age to assign the rights to a voice recording). For the presentation of this campaign, a local association with a large number of followers has been involved, the Galician video game museum (MUVI[15]). This will be a 100% digital campaign, which will try to spread through the collaboration of Galician streamers, creation of specific content and targeted advertising on social media. The great attraction of this campaign is the possibility of winning a prize. People who record a series of 30 sentences will be able to enter a draw for different prizes. Figure 9 shows the screen displayed to participants who complete the series of 30 sentences.



falAI Grazas por participar!

Parabéns, remataches!
Por ter gravado 30 frases podes participar no sorteo dunha PlayStation 5 ou outros premios!

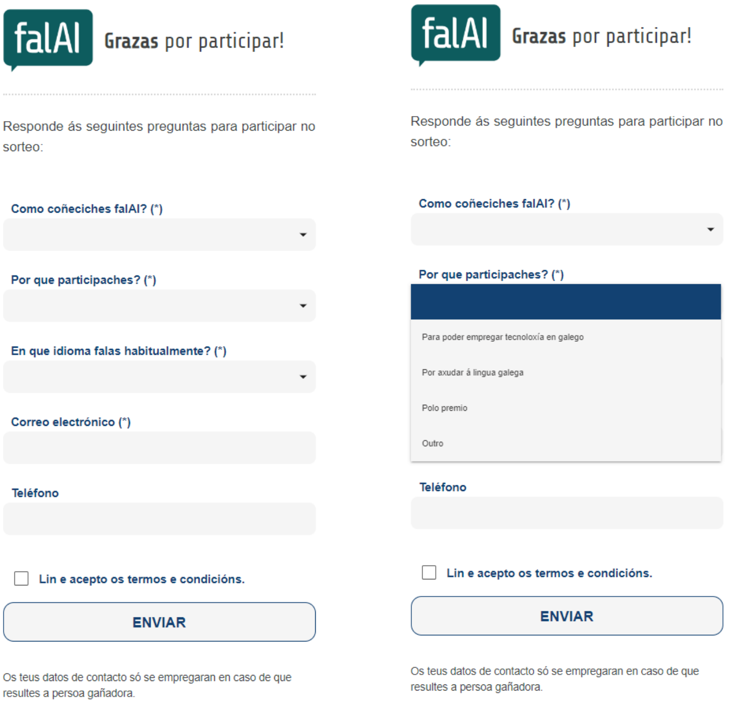
Se volves a gravar 30 frases poderás participar de novo no sorteo. **Cantas máis veces graves, máis posibilidades tes!**
Podes ver os datos da campaña en tempo real aquí.

Figure 9: Screen to enter the draw after completing the 30 sentences.

[14] The original one in Galician was: *a30frases*

[15] https://muvi.gal/

At this point the users can participate in the draw or repeat the series of 30 sentences. In case they want to take part in the draw, the next screens they will see are those shown in figure 10. In order to be able to measure which strategy is working best, to understand why people participate, and to better understand the donor profile, it has been decided to add three mandatory questions to this part of the challenge. We can allow ourselves to make them mandatory, something that would normally create negative reactions from participants, because only people who want to participate in the draw will reach this screen, i.e. people who are willing to do something else (in this case, answer three questions and share some contact information) in order to participate in the prize draw.



falAI Grazas por participar!

Responde ás seguintes preguntas para participar no sorteo:

Como coñeciches falAI? (*)

Por que participaches? (*)

En que idioma falas habitualmente? (*)

Correo electrónico (*)

Teléfono

□ Lin e acepto os termos e condicións.

ENVIAR

Os teus datos de contacto só se empregaran en caso de que resultes a persoa gañadora.

(a) Survey and participant data

falAI Grazas por participar!

Responde ás seguintes preguntas para participar no sorteo:

Como coñeciches falAI? (*)

Por que participaches? (*)

Para poder empregar tecnoloxía en galego

Por axudar á lingua galega

Polo premio

Outro

Teléfono

□ Lin e acepto os termos e condicións.

ENVIAR

Os teus datos de contacto só se empregaran en caso de que resultes a persoa gañadora.

(b) Survey drop-down example

Figure 10: Survey and participant data for the draw.

The questions in the final form are:

- How did you discover falAI?
  - Internet or social networks
  - Through another person
  - Through the press or television
  - Other
- Why did you participate?
  - To be able to use technology in Galician
  - To help the language
  - Because of the prize
  - Other

- What language do you usually speak?
  - Galician
  - Spanish

After answering these three questions, it will be necessary to provide at least one contact information (email or telephone) and accept the terms and conditions of the prize draw. This document[16] clearly states the dates of the draw, the mechanics, possible prizes and data protection. The contact details will never be used for any other purpose than to communicate with the winner (no commercial or promotional communication will ever be made).

To participate, they will be asked to accept the terms and conditions of the draw, provide contact details and answer three questions related to the reason for participating in falAI. Participation in the prize draw is optional and can be entered as many times as you want. For each series of 30 recorded sentences, one entry will be added to the prize draw. Both the initial launch event with streamers and the opportunity to receive the prize will be promoted with targeted advertising on social media.

4.3 Older people campaign

Campaign oriented to the participation and collection of voices of older people. We are aware of the difficulties that a non-technological profile may face when trying to access tools such as falAI, o strategies have been designed to get them involved. Part of the campaign will be will target additional groups of people (young people and even minors), who can help their elders to access the platform and record their elders. This will be accompanied by motivational[17][18] videos that will have an impact on social media and television. In addition, voice collection sessions have been planned in collaboration with Galician care homes, which can be used for media impact.

5 FalAI evolution

This section shows the progress of the campaign since its launch and details the different actions carried out as part of each of the strategies designed. Figure 11 shows the evolution of falAI in number of recordings per day since its launch until 31 March 2023.

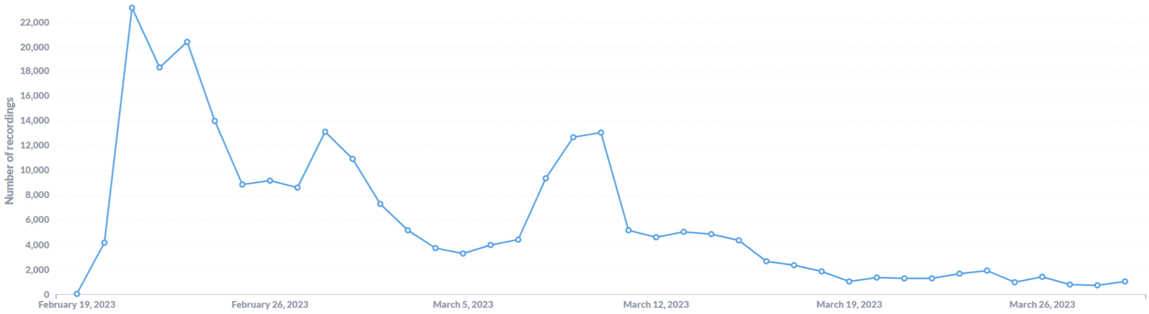


Figure 11: Evolution of falAI: number of recordings per day during the campaign.

[16] Accessible via: https://falai.balidea.com/bases_legais_sorteo_falai_gl.pdf

[17] **Video 1 accessible via: https://youtu.be/gE8E-yHMveE or https://www.instagram.com/p/CpcaxvKJxy8/**

[18] **Video 2 accesible via: https://youtu.be/wJtKf-xR43A or https://www.instagram.com/p/Cpca_JZOovR/**

5.1 Official launch

The official launch of falAI took place on 20 February 2023 in an event organised by the University of Vigo and led by the Vice-Chancellor for Communication and Institutional Relations, to which Balidea and the Mayor of Vigo representing the City Council of Vigo were invited. It was decided that the event would be organised and led by the University of Vigo in order to expose, at least to Vigo City Council and the media, that it is clearly a research project with a public purpose, where thanks to a public-private partnership in collaboration with Balidea it is possible to carry it out.

The strategy of further exposing and presenting the public organisation first and then the private company has been totally premeditated and organised. We are aware of the possible resistance to actions in which personal data (such as voice) is requested from the population in an altruistic and massive way. In many sectors of the population, there is a clear rejection and a high level of concern towards the data management currently carried out by companies and governments. It is for this reason that we decided to highlight the public part within the collaboration, as we believe that a project led by a public university can have a greater acceptance than a project led by a private company. Of course, the work of Balidea and the ELE project was also well explained in each of the public participations.

Vigo is the largest and most populated city in Galicia and is also where the research group (GTM) of the University of Vigo is located. In addition, the Mayor of Vigo, Abel Caballero, a very famous and locally relevant person, has the capacity to gather media on a massive scale and to create relevant impacts. It is for these reasons that the collaboration of the local government of the city where the University of Vigo is present was requested. In the presentation and in the information given to the media, it was explained in detail that Vigo City Council participates as a promoter, supporting the project and the collaboration, but making it clear that it is not a project of Vigo City Council, nor has it participated in its financing. Some examples of the many articles from that day and the following days include (eur, 2023), (met, 2023), (Fernández, 2023) or (Lado Alvela, 2023).

5.2 Youth Campaign Launch

Due to the time constraints of the project and the dependency of the organisations involved in the dissemination of the initiative, the official launch of falAI had to be delayed, which meant that it coincided with the launch of falAI as part of the young people's campaign. The launch of the campaign also took place on 20 February. As mentioned in section 4.2, this event was organised in collaboration with MUVI, an organisation with a large number of local followers and a clear sensitivity to the defence of the language and the creation of content in Galician. To this end, the Galician Video Game Museum itself organised a talk between representatives of Balidea, the University of Vigo and the MUVI, which would be moderated and broadcast live by two local streamers via Twitch, Facebook and Youtube[19].

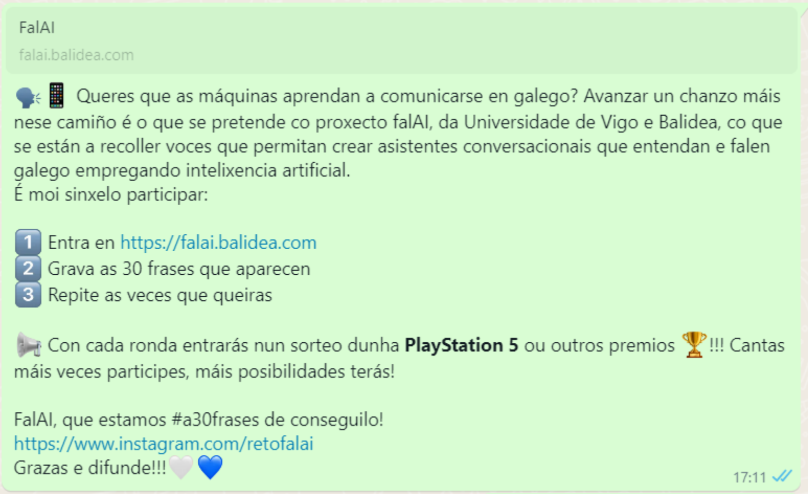
The aim of the talk was to present the project and the falAI tool in a distended way, as well as the attractions of the project, in terms of research but also in terms of the possible prizes of the draw, related to the video game industry. In this talk it was explained who the participants in the project are, where the project comes from, and the possibilities and benefits of this type of technology.

5.3 First reactions and campaign reception

In addition to the two presentation events, Balidea's communication team designed a message with all the information about the project (what is falAI, who participates, what it is

[19] Video of the talk accessible via: https://www.youtube.com/watch?v=oWv9EjSyU0M

for, how to participate and the possible prizes) to be spread individually together with one of the designed videos through instant messaging applications (in Galicia, the most used is WhatsApp) and to seek word of mouth to viralise the initiative. Figure 12 shows how the message was designed and disseminated.



FalAI
falai.balidea.com

🗣️📱 Queres que as máquinas aprendan a comunicarse en galego? Avanzar un chanzo máis nese camiño é o que se pretende co proxecto falAI, da Universidade de Vigo e Balidea, co que se están a recoller voces que permitan crear asistentes conversacionais que entendan e falen galego empregando intelixencia artificial.
É moi sinxelo participar:

1️⃣ Entra en https://falai.balidea.com
2️⃣ Grava as 30 frases que aparecen
3️⃣ Repite as veces que queiras

📢 Con cada ronda entrarás nun sorteo dunha **PlayStation 5** ou outros premios 🏆!!! Cantas máis veces participes, máis posibilidades terás!

FalAI, que estamos #a30frases de conseguilo!
https://www.instagram.com/retofalai
Grazas e difunde!!!🤍💙

17:11 ✓✓

Figure 12: Message designed to broadcast via instant messaging.

As can be seen in the figure 11, the reception of the campaign was a success from the beginning, resulting in more than 80,000 recordings in the first five days, and exceeding 100 hours of recordings (the original goal) in the first week. The reception in the media, mainly as a result of the presentation with the city council of Vigo, was massive, giving visibility to the project in Galician and Spanish newspapers, on local and regional radio stations, and on Galician Television, the most watched television channel in Galicia.

Moreover, we were able to verify that the word-of-mouth effect exceeded all our expectations, spreading the project throughout Galicia (to areas where the Vigo media could not reach in the same way) and very quickly. Sometimes, we found that when we tried to spread the message to different collectives, groups or areas, it had already arrived previously through another branch of dissemination (out of our reach).

The project and the proposal was also very well received in terms of image. The proposal was perfectly understood by the population, who liked the initiative and did not hesitate to share it altruistically, and Galician personalities quickly joined in and called for public collaboration in the project.

In social networks, the impact of the project was also high, although not yet through the project's official profiles. One of the most massive movements we could observe was on Twitter, where people not related to the project shared the idea and requested collaboration. Motivated by this unexpected effect, it was decided to present the project through this social network, but this time in the personal title of the main researcher. The idea is to also measure the reaction when the people involved in a project are personified, and to measure the possible impact through social networks. The specific tweet[20], published on 23 February 2023 (three days after the presentation) along with one of the promotional videos, achieved a high circulation, getting more than 55,000 views of the tweet and more than 12,500 views of the promotional video, with more than 400 retweets, more than 40 mentions and multiple comments and private messages. The tweet was retweeted by Galician public figures (from musicians to intellectuals) with thousands of followers on the social network. In terms of

[20] https://twitter.com/DoctorPinheiro/status/1628726875370528768?s=20

recordings, there was a clear spike once the tweet started to go viral (second peak of figure 11, more than 20,000 recordings). Figure 13 shows the hourly evolution in number of recordings after the tweet was published.

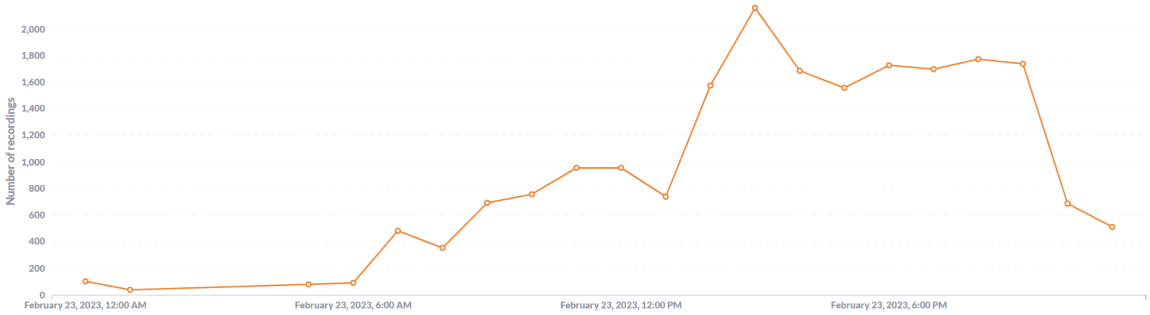


Figure 13: Number of recordings per hour after tweet viralisation.

5.4 Midterm analysis and older people campaign evolution

After the successful launch of the campaign, digital content continued to be created on the project's social media profiles, promoting competition between different cities, regions or accents, and seeking media impact (radio interviews and collaborations on Galician television programmes). It is at this point that it was decided to increase the corpus designed to take advantage of the success of the campaigns and achieve recordings in wider domains and with greater complexity. It is also at this point that it is found that many of the recordings are read incorrectly as they are made literally, so it is decided to introduce the written variant, which is correct orally (see section 3.3).

The next campaign strategy started in the third and fourth week of data collection. The aim of this campaign was to collect voices of older people, and the first action was a collaboration with language-related influencers to ask for participation in the campaign and to ask young people to record older people. In parallel, a video was also launched to show the situation of older people who do not have access to technology because of their language. Finally, a voice collection day was also organised in a residence on 23 March 2023. The aim of the day, beyond the voices that could be collected in the residence, was to attract media and create an impact in areas that had not been so widely involved to date. The day was very positive in terms of sensations with the older people who decided to participate, and local media also attended to cover the day (Reboredo, 2023).

5.5 FalAI results

From the very beginning, it was decided that the duration of the falAI project would not be strictly linked to the duration of the ELE project, in order to make the best use of the effort and resources available. It is for this reason that the results presented in this subsection are partial results, achieved as of 31 March 2023, the date on which this report is delivered. FalAI will remain active until at least the end of May, when the draw and prize-giving for participants will take place. Table 4 shows the main results achieved within the FalAI campaign:

The initial goal of 100 hours has been surpassed, an unprecedented milestone for the Galician, and we are close to surpassing the second goal of 300 hours. The number of recordings

Number of hours	250
Number of recordings	245,000
Number of participants	11,300
Municipalities participating	98.7%
Female / Male ratio	61% - 39%
Participants over 50 years	2,500
Hours from participants over 50 years	54.15

Table 4: FalAI main results.

has also far exceeded the initial objectives, achieving a total of 70 recordings per sentence. With regard to the number of participants, we can estimate that more than 10,000 people have participated in the falAI challenge and recorded their voice (it is not possible to draw a direct relationship between falai users and participants as it is possible that many people participated more than once). This data makes falAI, to the best of our knowledge, the largest database in terms of recordings, hours and participants publicly available for SLU, above those available in English such as STOP or SLURP. Finally, with regard to the gender of the participants, there is a clear tendency for women to participate and donate their voice in the majority, achieving more than 60% of the recordings.

The participation of the Galician population covered the entire territory of Galicia. As expected, the municipalities with more contributions have been the big cities (Vigo, A Coruña, Santiago de Compostela, Ourense, Lugo and Pontevedra), but a wide participation has been observed in less populated municipalities (Redondela or Ponteareas). Figure 14 shows a map with the participation ratio per municipality, where only 4 municipalities have remained without participation. If we look at the participation per municipality, 81% of the municipalities have had a significant participation of 5 or more participants (256 out of 313), and 50% of the Galician municipalities have had a participation of more than 15 people (157 out of 313).

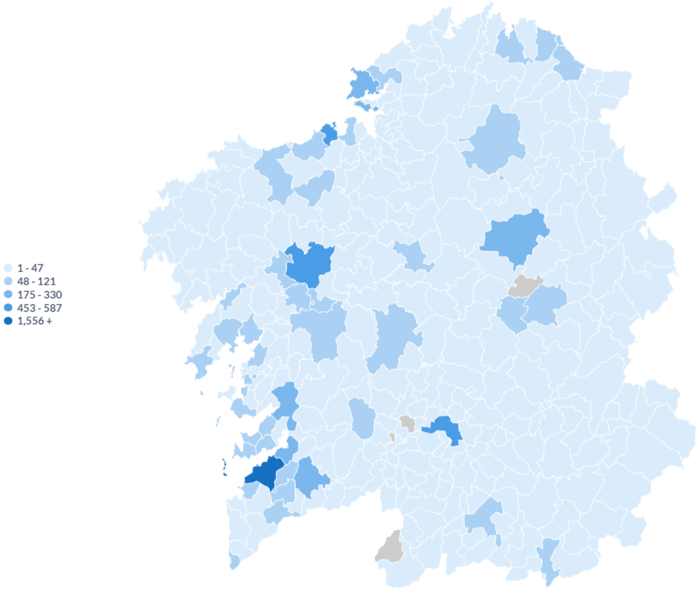


Figure 14: Participants per municipality.

If we analyse the results according to the provinces, table 5 shows the hours and number of recordings for each of the provinces, and figure 15 shows a graph with the number of participants per province. As expected, the most populated provinces have achieved the highest representation:

	Hours	Number of recordings
From A Coruña	83.48	81,784
From Lugo	35.24	34,146
From Ourense	27.05	26,212
From Pontevedra	97.11	102,349

Table 5: FalAI province results in terms of hours and recordings.

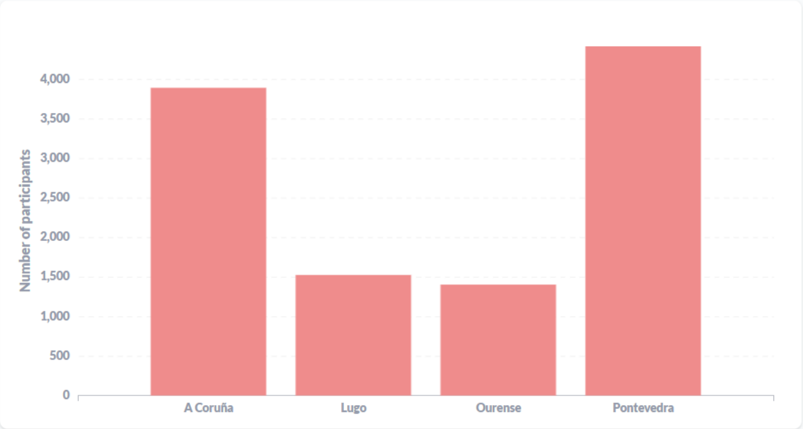


Figure 15: Participants per province.

FalAI has also been a success in terms of representation of each of the main variants of Galician (something that could be sensed on the basis of the participation throughout the territory). Table 6 shows the number of hours and recordings for each of the accents, where it is notable the number of participants who did not share their accent (probably because they were not sure which accent to use or because they did not clearly identify with any of them), and Figure 16 shows a graph with the distribution of participants among the different accents. The representation obtained matches the distribution of current speakers in Galicia (a higher number of normative speakers versus a lower number of speakers with an oriental accent, the least populated area of Galicia).

The representation obtained for each of the language variants will be of great interest for future socio-linguistic studies. Throughout the advertising of the campaigns, it has been emphasised that the participants have to use their natural speech, as if they were talking to a person, using their own accent variants and even changing words if they use other variants. After random validations, it was confirmed that most of the participants use the variants of Galician and clearly use one of the 4 main accents defined in Galicia.

FalAI has had a clear motivation to obtain voices of older people, typically groups that are under-represented in the available voice datasets. In natural speech applications, this under-representation causes systems to perform worse on these population groups. As shown in Table 7, thousands of people over the age of 50 have participated. The campaign has particularly engaged people between 30 and 50 years old. We believe that this group of people has particularly connected with the campaign, motivated by the need to help language and

	Hours	Number of recordings
Atlantic	31.52	30,668
Central	30.16	29,551
Neo-speaker	25.32	24,808
Normative	101.90	99,547
Oriental	21.05	20,501
Not-shared	38.98	39,564

Table 6: FalAI accent results.

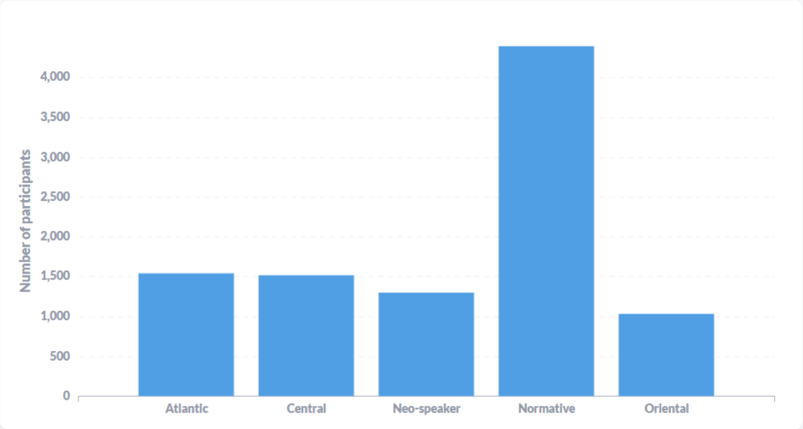


Figure 16: Participants per accent.

people who need to use this technology. Figure 17 shows the graph with the distribution of participants by age range.

	Hours	Number of recordings
Under 19	19.75	19,537
Between 20 and 29	44.48	44,488
Between 30 and 39	53.21	52,411
Between 40 and 49	76.69	76,909
Between 50 and 59	36.17	34,876
Between 60 and 69	14.35	12,764
Between 70 and 79	3.21	2,665
Over 80	0.7	449

Table 7: FalAI age range results.

6 Validation and analysis over the collected dataset

Validation of a speech file dataset can be a complex process that requires a combination of different strategies. Due to the validation control in place to avoid ghost recordings or recordings different from the reference sentence, we know that all recordings have a minimum quality and that they resemble the expected sentence.

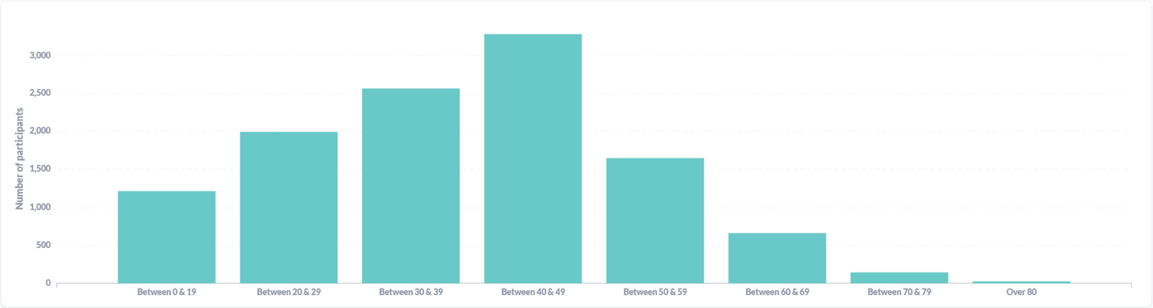


Figure 17: Participants per age range.

If we review the validation techniques used in other datasets, we see that SLURP (Bastianelli et al., 2020) uses a fully automatic validation, where they first evaluate with two ASRs if they get a perfect transcription (60% of their dataset), and then they relax the threshold to see if the entity fills are perfect (EntityWER=0), managing to automatically validate 73% of their dataset. Regarding the STOP dataset (Tomasello et al., 2023), they perform a semi-automatic validation with an ASR where, if the CER is below 50%, they add the recording directly to the dataset. For recordings below this threshold, they perform a manual validation. Neither of these two datasets add additional metadata in the validations.

In this section we will explain the validation strategy we have designed, taking into account that it is not the same to validate a dataset in English, where there are ASRs trained with thousands of hours of audio, than in a language with few resources such as Galician. The validation strategy presented can be extrapolated to languages with a similar situation to Galician.

6.1 Semi-automatic validation

Because the ASRs at our disposal do not have a confidence comparable to that of English, because in Galician, changes in a single character can modify the meaning of a sentence, and because the complexity of some of the sentences is high, including numbers, names of municipalities or acronyms, the validation designed for the falAI dataset will be a semi-automatic validation. It should be noted that within the ELE project it has not been possible to validate the achieved dataset (more than 250 hours with a margin of weeks), but the definitive validated dataset will be made public in the coming months.

For semi-automatic validation, a fine-tuned XLS-R (Babu et al., 2021), a large-scale model for cross-lingual speech representation based on wav2vec 2.0 (Baevski et al., 2020), has been used as the ASR model. This model, available in the HuggingFace library (Wolf et al., 2020), has been trained with 128 languages, including Galician and many European languages, and allows obtaining acceptable speech recognition results with a few hours of audio (in the case of Galician, with 12 hours of audio it is possible to train perfectly valid models). The validation strategy is first to normalise the audio, eliminate the silences at the beginning and end of the audio automatically, and then calculate the transcription with the XLS-R ASR model and, in case of WER=0, automatically validate the recording. If a perfect transcription is not achieved, the recording will be validated manually.

For manual validation, it has been decided to introduce metadata associated with each recording, in order to obtain a richer, more interesting and useful validated dataset for future studies. From the SLU perspective, one does not necessarily need the person to say exactly the reference sentence in order to have the intention and entities correct, i.e., as long as the recording carries the information to be understood, it will be correct, even if he/she changes some words, adds or removes words, or hesitates in its pronunciation. For this reason it has been decided to add the metadata present in table 8. Automatically validated recordings will be labelled as “validated” (assuming that the noise level is acceptable and that nothing has been changed in the reference sentence).

Metadata	Meaning
Validated	Utter which corresponds exactly with the reference sentence.
Change	Utter in which some word/s have been changed or pronounced differently, but retains the information of intent and entity
More words	Utter in which words are added, but which retains the information of intention and entity.
Less words	Utter in which words are omitted, but which retains information on intent and entity
Hesitation	Utter in which there is hesitation in pronunciation, whether or not the reference sentence is changed, but the semantic information is retained
Noisy	Utter noisy, where the reference sentence changes or not, but the semantic information is retained
Other	For any other case

Table 8: Metadata labelling in manual validation.

For automatic validation, a strategy has been devised that is divided into several phases. In a first phase, the recordings are randomly reviewed and validated with the available ASR and manually. When a considerable number of validated recordings are available (in our case, 10,000 recordings), a fine-tuning of the ASR model with the validated sentences (automatically and manually) is proposed, completely adjusting its performance to the domain of the dataset to be recognised. Moreover, it is possible to further boost inference if we use a language model trained specifically for this task. We have tested creating an n-gram language model (in our case, pentagrams) with the dataset text corpus and with Galician Wikipedia text corpus, and by incorporating the 5-gram model in the ASR decoder the recognition results improve by reducing the WER by 5 percentage points. For languages with few resources, this strategy can considerably increase the ratio of automatically validated recordings, reducing the manual effort required. Of course, it is possible to repeat this process with more validated sentences.

In addition, this labelling strategy allows the validated dataset to be used for multiple purposes. The recordings labelled as “validated” can be used to train ASR models (as has been done in the second phase of the automatic validation strategy), those labelled as “change”, “less words” or “more words” allow to train or test the future SLU model in scenarios more similar to a real scenario, and those labelled as “noisy” or “hesitation” allow to train or test our system in more complicated situations (studies not yet contemplated for E2E SLU).

6.2 Quality metrics over collected dataset

As discussed in the falAI results section, the audio data achieved in falAI is spectacular, surpassing the main and largest publicly available SLU datasets, to the best of our knowledge, in terms of hours, recordings and speakers. The success is especially remarkable in terms of the number of speakers achieved, with representations of all age ranges (except children) and variants of the Galician language. Table 9 shows a comparison between the main datasets.

With respect to audio quality, we know that due to the nature of the recordings, the audio quality is lower than in datasets where the recordings have been made in controlled envi-

	FSC	SNIPS	SLURP	STOP	**falAI**
Speakers	97	67	177	885	**11,300**
Audio files	30,043	5,886	72,277	236,477	**245,000**
Duration [hrs]	19	5.5	58	218	**250**

Table 9: Spoken dataset comparison.

ronments. In addition, problems have been detected with the recording platform for some specific devices or very old devices[21]. However, the variety of speakers and acoustic environments present in the dataset makes it worth losing some quality in the audio recording. Even so, we have been able to verify that the quality of the recordings is above expectations, finding a very low percentage of noisy or incorrect recordings. Table 10 shows the main quality measures proposed for an SLU (or speech) dataset:

Number of hours	250
Number of files	245,000
Speakers	11,300
Average SNR [dB]	33.07[22]
Annotation	Yes
Metadata labelling	Yes[23]
Acoustic environment variety	High[24]

Table 10: Quality measurements over collected dataset.

7 Lessons learned

One of the objectives of the project was to extract conclusions from a use study, based on our experience, conclusions that could be useful for similar projects or for European languages that are in a similar situation to Galician, explaining what has worked, why, and how it would be possible to transfer the success to another language. This section outlines these lessons learned throughout the duration of the project.

7.1 Key points in the approach

In this sub-section, we have summarised what we believe to be the key elements in the approach and design of the initiative:

Transparent communication and a clear basis: We know that one of the keys to avoid people's rejection and to motivate not only their participation, but also their diffusion, has been the transparency in terms of the project's objectives, the use of voices, and the fact that it is "open" knowledge. The data release document, drawn up with legal experts, has been a

[21] Although the bug could not be fully addressed, we have detected problems with the library used for recording on some iOS devices and on devices whose limited resources were limiting the performance of the tool. Due to the time constraints of the project we have not had time to fix it and deploy a new version of the tool. We estimate that the affected recordings are below 5%.

[22] Calculated on 5,000 randomly selected recordings.

[23] Seven types of label.

[24] Proportional to the number of speakers.

key element to avoid possible criticism or problems with the use of voices, which is currently very sensitive, and to provide security. Throughout the campaign, in every interview or promotional action, the same message has been emphasised: the aim is to create technology in Galician, regardless of who does it. This project seeks to provide solutions. Analysing the feedback from the participants, we have seen that this message has been well received and has given peace of mind to participate and share. *A data collection project must be fully transparent regarding the purpose and processing of the data.*

Public-private collaboration: The collaboration with a public organisation such as the University of Vigo has been a key factor for the success of the project. With our experience in the project, we are convinced that this initiative would not have had the same impact if only a private company was behind the project. *This highlights that it is necessary for this type of initiative to come from public organisations, such as the government, or from non-political public knowledge organisations, such as universities or research institutes*. Of course, it is essential that there is coordination at national or regional level with similar public projects or governments, so that the effort made by such proposals is carried out efficiently, making the best use of resources and without duplicating work or hindering progress. In short, *the success of the public-private partnership has been based on leveraging the impact, image and knowledge of the public side and the implementation capacity and experience of the private side*.

Communicate an idea: We think that giving the project a commercial name (**falAI**), an attractive image and a simple idea that could be understood by any population profile has also been one of the reasons for the project's success. Being able to publicise the message "let's help machines understand Galician" or "let's create assistants who understand Galician" (messages that were widely shared during the campaign) meant that the population understood what we wanted to do and why. It is not always easy to convey technical information and to filter out what may not be so well understood by a general profile, so *we believe it is necessary to dedicate part of the effort to create an attractive and understandable "product" for the population, making it clear why it is necessary and what their effort will serve for*.

Linguistic content of the dataset: The linguistic content of the dataset itself has also helped to broaden the scope of the project, which was unexpected for us at the beginning. We have seen from the feedback on social media that many people joined in sharing falAI because of the phrases in the dataset. When designing the dataset, we decided to include references to places in Galicia, local festivals in Galicia, Galician music groups and artists, local expressions (even the most informal ones), Galician writers and books or Galician products, as well as sentences from many domains that we thought could be useful for the Galician population. This has been appreciated by many participants, who have decided to share how their part of Galicia was also represented, and it has also caused Galician artists or writers, with a large number of followers in social networks, to be mentioned and share the initiative. In short, the fact that participants found their own expressions, that they found the sentences useful, and that it was insisted that they use their own accent, their own words, their own language (often forgotten or under-represented), has helped to make it so well received. *People participated because they found it useful, hence the importance of designing application oriented.*

7.2 Key points for participation

This sub-section summarises what we believe have been the key actions to achieve the massive participation of the entire Galician population.

The recording tool: We have been able to see from the feedback of the participants how important it was that the tool was stable, simple and attractive. It is common that applications or tools developed by public organisations do not pay as much attention to this kind of details. In the design and creation of the tool, great care was taken to ensure that it was intuitive and simple for anyone, regardless of their technological profile, that it was accessible from any device and that it was attractive, not forgetting that the tool must also be able to handle recording peaks and that security measures must be implemented to avoid ghost recordings or denials of service. *A campaign like this must have a stable, multi-device, intuitive and well-maintained recording tool. It is not an option to lower the standards of design and performance at this point.*

The dashboard and the challenge idea: We believe that presenting the campaign as a challenge, which could be followed in real time, has been one of the main attractions in motivating participation. The videos and other digital content created highlighted the fact that it was a challenge (30 sentences away), and with the help of the dashboard in real time presented in section 4 we have been able to see how competitiveness arose between different areas of Galicia, how people, with the objective in mind, shared the initiative to ensure that the Galician language had the necessary resources to meet the challenge. The dashboard has meant that there was an image (a map, a graph) accessible to everyone and that it could be shared on social networks, instant messaging applications, etc. *The campaign has benefited enormously from having a clear motivational idea (the challenge) and a space where attractive campaign data can be viewed in real time (the dashboard).*

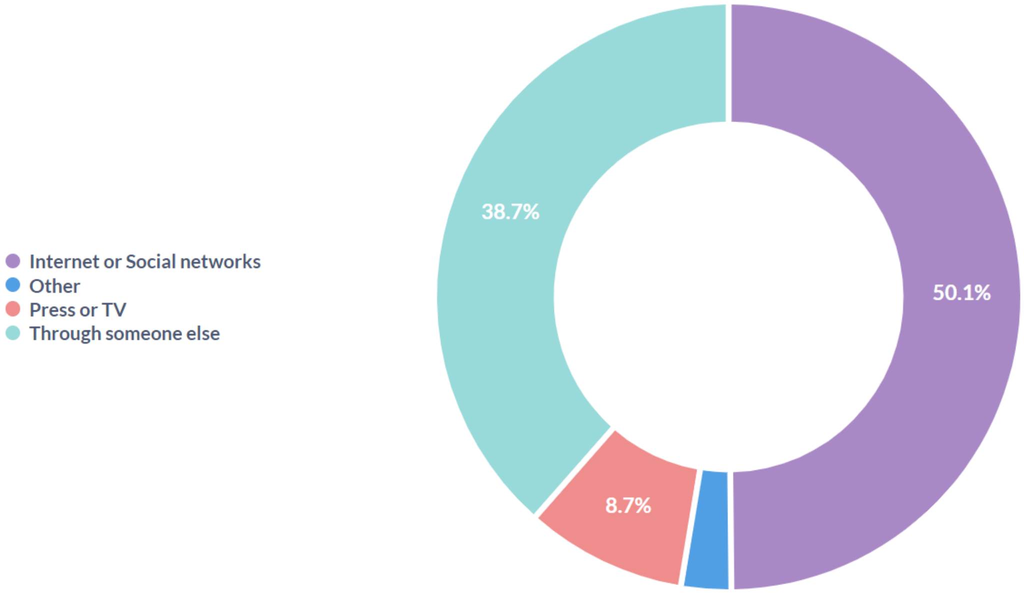


Figure 18: How people have known falAI.

Word-of-mouth: The word-of-mouth effect has exceeded all our expectations and, although it is difficult to measure correctly, we know that it has brought the initiative to areas or groups that had not yet been or were not going to be impacted by any kind of publicity action. Only word of mouth explains that hours after the launch in Vigo, the initiative was highly shared and we registered a high participation in areas far away from Vigo, such as Foz, Burela, Xinzo de Limia or Ares. We have been able to see on social networks, in forums and individually, how our actions to disseminate falAI were not the first falAI message to reach users. The WhatsApp message designed had a great impact (Figure 12), and we saw how this message or modifications to it spread massively. In addition, the voluntary responses of the users support these conclusions, since more than 38% say that they have known about falAI through another person (Figure 18). *Word-of-mouth has been the second most important disseminator of information. We believe it is necessary to focus efforts on facilitating this type of communication and creating seeds for its dissemination.*

Social networks and influencers impact: The impact achieved through social media was undoubtedly the most effective action to publicise falAI. We have been able to see how there has always been a peak after each of the collaborations with influencers or after some of the more viral actions. Figure 19 shows the campaign's two biggest peaks of recordings, thanks to collaborations with influencers (the 10 March peak achieved more than 3,000 recordings in less than an hour). However, it is difficult to predict which message or collaboration will have the greatest impact. There have even been unplanned promotional actions, where an influencer on TikTok with more than 1.8 million followers shared the proposal when she came across it on social media, causing a spike in recordings. If we look at the responses collected in the final form, more than 50% of the participants confirm that they met falAI through the internet or social networks. *In today's society, the communication element with the greatest potential is the internet and social networks. To publicise this type of campaign, it is necessary to have a presence on the internet and to collaborate with influencers and celebrities who are sensitive to language.*

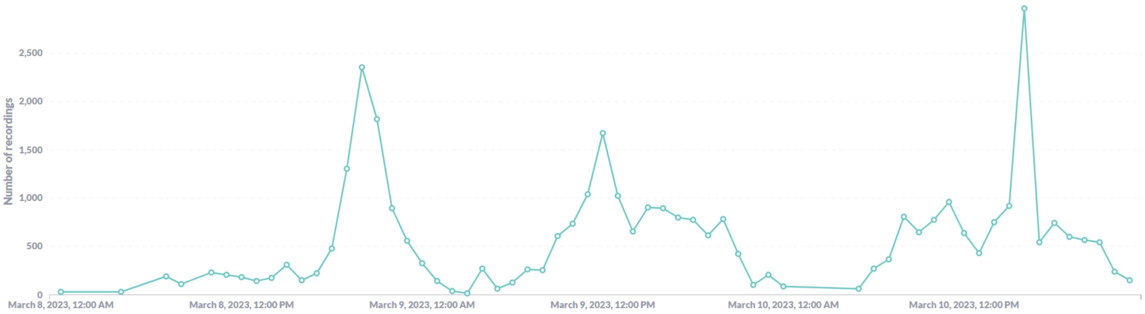


Figure 19: Number of recordings per hour after the collaboration of two influencers.

The reason: If the internet and social networks have been falAI's best advertising channel, the fact of being able to help the Galician language and feeling part of the proposal has undoubtedly been the main reason why people have decided to participate. All our actions emphasised this message: let's ensure, together, that Galician is also present in the digital world, and all the feedback we have received has been in line with this message. If we analyse the answers in the final form (figure 20), we see that 57.5% of the participants did it to help the language, and 37.7% participated because they want to be able to use technology in

Galician, that is, more than 95% of the people participated because they want to help or want to use the Galician language. *Getting people to feel involved in this proposal and to understand that it was necessary for the language has been the most important motivating element of the campaigns*.

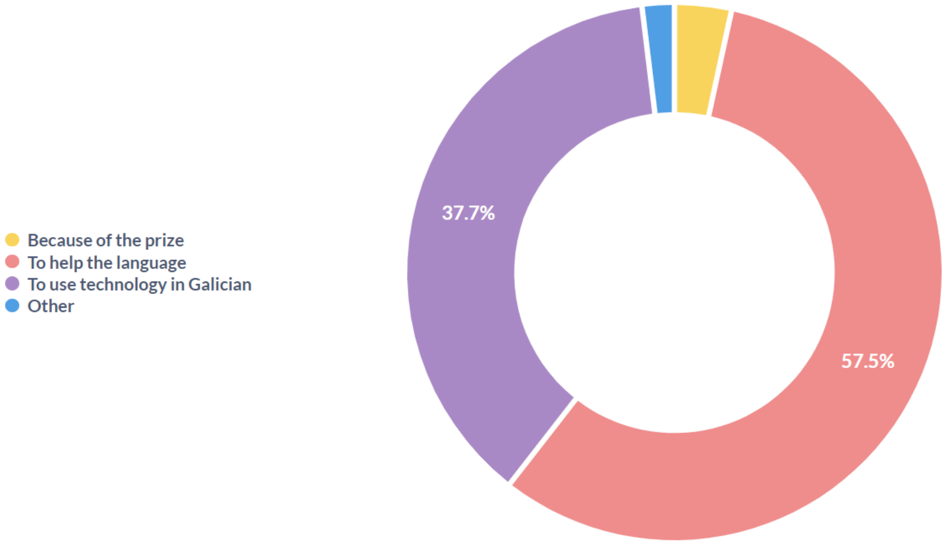


Figure 20: Reasons for participation in falAI provided by users.

Campaign design and programmed impacts: The correct planning of the campaign is key to its success. In our case, a good kick-off and presentation, accompanied by media, followed by social media promotion and digital content promotion worked well. Analysing the evolution of the recordings throughout the campaign (Figure 11), it is clear how the programmed impacts throughout the campaign are essential to maintain the level of participants. *The campaign must be well planned, with previously created and validated content, planned impacts and agreed collaborations to maximise its impact*.

Media presence: The presence in the media has been a great promoter of the initiative, especially in the official presentation of the project. The traditional media (press, radio, television) do not have the same impact as in the past, but the wide initial coverage has served to reach many people quickly and word of mouth has started to work. It has also given the project a sense of formality, allowing many people to join in and share, as they realised that it was an important project, with wide coverage in multiple media. *The presence in traditional media is an important action that should be complemented with the rest of the activities, helping to publicise the project at the launching and giving the initiative a strong packaging*.

The prize: We cannot ignore the fact that the prize was also a motivator for people to participate. Although the final form does not reflect this as the main motivator (less than 5% in figure 20), the feedback received internally, in groups and on social media does support that, even though not as the main motivator, the prize has been a complementary element that has helped participation in the challenge and in falAI. On the negative side, we believe that our strategy of offering several prizes to be chosen by the winner (in order to offer attractive prizes to a higher percentage of the population) has not been successful. We did not communicate it correctly, and most people thought there was only one prize. *In conclusion, for the majority of the population, the prize has not been the main motivating factor for participation, although it has helped to complement it, especially among younger participants*.

7.3 Actions where measuring the impact has not been possible

During the campaign, many actions have been carried out within the strategies set out but also on an improvised basis as opportunities arose during the evolution of the campaign. For many of these actions it has not been possible to measure impact or the impact has not been as expected. This sub-section presents the actions for which the feedback received does not allow any conclusions to be extracted.

Press impact on a stand-alone basis: Some of the appearances in the press (newspapers, radio or television) have been made on an isolated basis, once they have arisen. We have been able to see how these types of actions, in isolation, have zero or little impact. The last appearance in the press (23 March 2023) as a result of the collaboration with an elderly people's home was covered in the local press, but the recordings increased by 16% and only for a few days. The same has happened with similar actions throughout the campaign.

Advertising in isolation: Our isolated publicity actions in front of many people have also had little impact. The falAI advertising video was screened at an event with more than 3,000 people, with no change in trend over the next few days. Except with influencers, individual actions have had little impact. A continuous advertising exercise is necessary to achieve effects through this strategy.

Intergenerational collaboration: This strategy has not worked as we had hoped either, as we have not been able to measure a significant impact on this type of collaboration. The people who participated did so on their own, and we were not able to motivate young people to record their elders on a massive scale. However, due to time constraints and restrictions at Galician level, we have not been able to collaborate with schools and language standardisation organisations. Although we have not been able to test it, we believe that this type of strategy could work for this group.

Older people campaign: The specific strategy towards older people has not had the expected effect either. The project has been a success in terms of collecting the voices of people over 50 years old, but we have not measured an increase in recordings in this group with the specific actions of the campaign, but it was a trend since the launch of falAI. It should be noted that the actions of this campaign came weeks after the launch, where the surprise effect had already worn off. We believe that the data collection campaign would have been more effective if we had devoted more resources to on-site recordings in care homes throughout Galicia, where the initiative was well received by users and professionals.

Compromise between design and deployment: The time constraints of the project have affected the desired campaign design and implementation times. We believe that better planning and design would benefit future campaigns, where digital campaign content is available prior to launch, with a minimum of 3 months spent on campaign design and strategies. We believe that a future campaign should be launched with digital content and events already planned and validated, and with a powerful kick-off event (including replicating the event in different locations).

8 Conclusions

In this project we have designed a dataset for SLU from scratch, planned a data collection campaign, and presented measures that allow us to establish the complexity of similar text datasets, minimum design features, and measures on the quality of the dataset collected, regardless of the language of the dataset. The dataset designed is an unprecedented milestone for Galician, a low-resources language and with weak technological presence. The success of the campaign has been such that it has made it possible to **obtain the largest publicly available dataset for SLU**, to the best of our knowledge, above languages such as English, French or Chinese, with **more than 250 hours of recordings and more than 11,000 participants**.

The lessons learned in our campaigns, presented in the section 7, will serve as a guide for the design of similar campaigns in European languages in a similar situation to Galician, learning from our successes and our mistakes, allowing us to invest resources more efficiently. They show the importance of **good communication and transparency in the use of data**, the potential of **public-private collaboration**, the good results of investing resources in creating an **attractive idea**, or the importance of putting **technology at the service of society, designing people-oriented applications**. It is also exposed how **a simple but functional tool, real time data, word-of-mouth and mainly internet and social networks are key to the success of a similar campaign in the European society**.

The trend seems to indicate that chatbots and conversational assistants will be one of the main channels of communication with the digital world of the future, and E2E SLU technology may be one of the key elements for their operation. Advances in the creation of synthetic data are promising, increasing expectations that these types of technologies will become more prevalent. Even with the presence of synthetic data, the European languages will still need to have a minimum of real data for the creation of their dialogue systems. This project has presented guidelines and results based on the experience of a real, scalable and replicable project in multiple European scenarios and has focused on putting technology at the service of society, in line with the demands of today's European society and the ELE project.

8.1 Future directions

In the coming months, data will continue to be collected in the falAI tool and some of the strategies that were possible due to the time constraints of the project will be tested.

In Galicia, due to its situation of bilingualism, its historical context and the evolution of the number of Galician speakers, there is a network of language promotion and standardization teams coordinated by the Galician government (*Secretaría Xeral de Política Lingüística*). The objectives of this network are, among others, to promote the Galician language, to ensure institutional contact and to broaden the dissemination of the actions of the different entities that collaborate with the network. In turn, the network and teams collaborate closely with the Galician Department of Education, coordinating actions also in educational centers throughout the country.

We believe that a great strategy would be to establish a framework of governmental collaboration throughout Galicia between the linguistic normalisation teams and the educational centres, with the aim of presenting the project on a massive scale in the educational centres. During the duration of the project it has not been possible to carry out these actions due to the time constraints of the Galician government, as these types of actions have to be validated and coordinated with the Ministry of Education, a process that can take months and involves considerable effort.

Finally, once all data collection campaigns have been completed, a review and validation of the collected dataset will be carried out. After this process, access to the collected data will be made publicly available. The collected and validated data will be used in real projects to create conversational assistants in Galician by Balidea and the University of Vigo, and the results of their performance will be published when available.

References

Uvigo y balidea impulsan un proyecto para recoger voces y ayudar a la tecnología a hablar y entender el gallego. *Europa Press*, February 2023. URL https://www.europapress.es/galicia/digital-01089/noticia-uvigo-balidea-impulsan-proyecto-recoger-voces-ayudar-tecnologia-hablar-entender-gallego-20230220184409.html.

A uvigo e balidea queren que os asistentes por voz falen galego e precisan de ti. *Metropolitano.gal*, February 2023. URL https://metropolitano.gal/enfoque/a-uvigo-e-balidea-queren-que-os-asistentes-por-voz-falen-galego-e-precisan-de-ti/.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. Slurp: A spoken language understanding resource package. *arXiv preprint arXiv:2011.13205*, 2020.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo Lopez-Gazpio, and Lucia Specia. Semeval-2017 task 1: Semantic textual similarity-multilingual and cross-lingual focused evaluation. *arXiv preprint arXiv:1708.00055*, 2017.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*, 2018.

Iria de Dios-Flores, Carmen Magarinos, Adina Ioana Vladu, John E Ortega, José Ramom Pichel Campos, Marcos Garcia, Pablo Gamallo, Elisa Fernández Rei, Alberto Bugarín Diz, Manuel González González, et al. The nós project: Opening routes for the galician language in the field of language technologies. In *Proceedings of the Workshop Towards Digital Language Equality within the 13th Language Resources and Evaluation Conference*, pages 52–61, 2022.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

ELE-Consortium. Deliverable D3.4 Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/11/ELE___Deliverable_D3_4__SRIIA_and_Roadmap___final_version_-1.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*, 2020.

Sergio Fernández. La universidad de vigo busca voces para que siri y alexa hablen en gallego. *Atlántico*, February 2023. URL https://www.atlantico.net/articulo/universidad/universidad-vigo-busca-voces-que-siri-alexa-hablen-gallego/20230220230034969451.html.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. Semantic parsing for task oriented dialog using hierarchical representations. *arXiv preprint arXiv:1810.07942*, 2018.

Parisa Haghani, Arun Narayanan, Michiel Bacchiani, Galen Chuang, Neeraj Gaur, Pedro Moreno, Rohit Prabhavalkar, Zhongdi Qu, and Austin Waters. From audio to semantics: Approaches to end-to-end spoken language understanding. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 720–726. IEEE, 2018.

Charles T Hemphill, John J Godfrey, and George R Doddington. The atis spoken language systems pilot corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*, 1990.

I.G.E. Enquisa estrutural a fogares. coñecemento e uso do galego. resumo de resultado 27/09/2019. sep 2019a. URL https://www.ige.gal/estatico/estatRM.jsp?c=0206004&ruta=html/gl/OperacionsEstruturais/Resumo_resultados_EEF_Galego.html.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. 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 2020. European Language Resources association (ELRA). ISBN 979-10-95546-35-1. URL https://www.aclweb.org/anthology/2020.sltu-1.3.

Juan Ventura Lado Alvela. A universidade de vigo busca voluntarios para ensinarlle á intelixencia artificial a falar galego. *Voz de Galicia*, February 2023. URL https://www.lavozdegalicia.es/noticia/sociedad/2023/02/21/span-langgl-universidade-vigo-busca-voluntarios-ensinarlle-a-intelixencia-artificial-falar-galegospan/0003_202302G21P23992.htm.

Batia Laufer. The lexical profile of second language writing: Does it change over time? *RELC journal*, 25(2):21–33, 1994.

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. Speech model pre-training for end-to-end spoken language understanding. *arXiv preprint arXiv:1904.03670*, 2019.

Joseph P McKenna, Samridhi Choudhary, Michael Saxon, Grant P Strimel, and Athanasios Mouchtaris. Semantic complexity in end-to-end spoken language understanding. *arXiv preprint arXiv:2008.02858*, 2020.

Bernadette O'Rourke. The galician language in the twenty-first century. *A companion to Galician culture*, pages 73–92, 2014.

Arturo Reboredo. Os maiores aprenden á intelixencia artificial a falar galego. *El Progreso*, March 2023. URL https://www.elprogreso.es/articulo/comarcas/maiores-aprenden-siri-falar-galego/202303241233361651144.html.

Dmitriy Serdyuk, Yongqiang Wang, Christian Fuegen, Anuj Kumar, Baiyang Liu, and Yoshua Bengio. Towards end-to-end spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5754–5758. IEEE, 2018.

José Manuel Ramírez Sánchez and Carmen García Mateo. Deliverable D1.15 Report on the Galician Language, 2022. URL https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_15__Language_Report_Galician_.pdf. Project deliverable; EU project European Language Equality (ELE); Grant Agreement no. LC-01641480 – 101018166 ELE.

José Manuel Ramírez Sánchez, Laura Docio-Fernandez, and Carmen Garcia Mateo. Galician's language technologies in the digital age. In *Proc. IberSPEECH 2022*, pages 21–25, 2022. doi: 10.21437/IberSPEECH.2022-5.

Paden Tomasello, Akshat Shrivastava, Daniel Lazar, Po-Chun Hsu, Duc Le, Adithya Sagar, Ali Elkahky, Jade Copet, Wei-Ning Hsu, Yossi Adi, et al. Stop: A dataset for spoken task oriented semantic parsing. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 991–998. IEEE, 2023.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020.