FalAI: A Dataset for End-to-end Spoken Language Understanding in a Low-Resource Scenario

Andrés Piñeiro-Martín[1,2], Carmen García-Mateo[1], Laura Docío-Fernández[1], María del Carmen López-Pérez[2], José Gandarela-Rodríguez[2]

[1]GTM research group, AtlanTTic Research Center, University of Vigo, Vigo, Spain
[2]Balidea Consulting & Programming S.L., Santiago de Compostela, Spain
{andres.pineiro, carmen.lopez, jose.gandarela}@balidea.com,
carmen.garcia@uvigo.gal, ldocio@gts.uvigo.es

Abstract

End-to-end (E2E) Spoken Language Understanding (SLU) systems infer structured information directly from the speech signal using a single model. Due to the success of virtual assistants and the increasing demand for speech interfaces, these architectures are being actively researched for their potential to improve system performance by exploiting acoustic information and avoiding the cascading errors of traditional architectures. However, these systems require large amounts of specific, well-labelled speech data for training, which is expensive to obtain even in English, where the number of public audio datasets for SLU is limited. In this paper, we release the FalAI dataset, the largest public SLU dataset in terms of hours (250 hours), recordings (260,000) and participants (over 10,000), which is also the first SLU dataset in Galician and the first to be obtained in a low-resource scenario. Furthermore, we present new measures of complexity for the text corpora, the strategies followed for the design, collection and validation of the dataset, and we define splits for noisy audio, hesitant audio and audio where the sentence has changed but the structured information is preserved. These novel splits provide a unique resource for testing SLU systems in challenging, real-world scenarios.

Keywords: spoken language understanding, end-to-end, SLU datasets

1. Introduction

Over the past decade, there has been growing interest in the development of voice-based virtual assistants (VAs), not only from technology companies, but also from governments and administrations for use in public services. Recent breakthroughs in Natural Language Understanding (NLU), Automatic Speech Recognition (ASR) and Large Language Models (LLMs) have made it possible to interact with machines in a more natural and fluent way in a wider range of contexts, thereby normalising voice-based interactions with the virtual world. The demand for speech interfaces is growing and is expected to increase exponentially in the coming years as areas such as education (Kasneci et al., 2023; Gubareva and Lopes, 2020) or health (Vona et al., 2020) become more important.

Spoken Language Understanding (SLU) is an interdisciplinary field that aims to extract structured information from speech signals, and lies at the intersection of speech recognition and natural language understanding. SLU systems are critical components of virtual assistants and are traditionally designed as a pipeline with an ASR module followed by an NLU module. However, these traditional architectures have several drawbacks, including the fact that their components are optimised according to different criteria, they suffer from cascading errors, and they waste computational resources. End-to-end (E2E) architectures have been successfully used in areas such as machine translation, speech synthesis or automatic speech recognition, and in recent years E2E architectures for SLU have been an active area of research (Serdyuk et al., 2018; Haghani et al., 2018). In these systems, the structured information is extracted directly from the speech signal and it is possible to leverage acoustic information such as prosody. The main research challenge in these architectures is the lack of large and challenging speech datasets with structured information or semantic parsing labels, and the low semantic complexity and limited domain coverage. The challenge is even greater for languages other than English, where public datasets exist only for Mandarin Chinese (Zhu et al., 2019), Indian (Rajaa et al., 2022) or Italian (Bellomaria et al., 2019).

In this paper we present the FalAI[1] dataset, which is the largest publicly available dataset for E2E SLU in terms of hours (over 250 hours of speech-labelled data), recordings (over 260,000 recordings) and participants (over 10,000 participants, 12 times more than the largest public dataset to

[1]Composition between the Galician verb for speaking, "falar", and the acronym for artificial intelligence, AI. Also, in some areas of Galicia, this is how the second person plural of the imperative of the verb to speak is formed: "falai" instead of "falade".

LREC-COLING 2024, pages 7107–7116
20-25 May, 2024.

date). The dataset covers 14 domains, 62 intents and 64 slot types. In addition, FalAI is the first publicly available SLU dataset in Galician, generally considered a low-resource language with large linguistic variation, making it an unprecedented milestone for the language. The dataset was designed, acquired and validated during the first semester of 2023, and the methodology and strategies followed are also presented in this paper. In addition, this paper introduces new complexity measures for text corpus design and presents splits for noisy audio, audio with hesitations, or audio with transcripts other than the reference sentence but preserving the structured information in the form of domain, intent, and slots. These novel splits will allow the establishment of a benchmark to test SLU models in borderline scenarios (noisy environments) or more realistic scenarios (audios with hesitation), as well as to test the adaptability and generalisation of the models (modifications of the original sentence with the same structured information, dialectal variations of the language, accent variations, etc.).

The rest of the paper is organised as follows: Section 2 outlines the related work on SLU datasets and the Galician context for language technology. Section 3 introduces the dataset design, collection and validation. Section 4 describes the lexical and semantic analysis of the dataset and presents the complexity measures. Finally, Section 5 contains the results and analysis, and Section 6 the conclusions and future directions of the work.

2. Related Work

The first speech dataset in the literature to include audio and annotated structured information was the Air Travel Information System (ATIS) (Hemphill et al., 1990), introduced in the 1990s. The success of end-to-end architectures in areas such as machine translation (Sutskever et al., 2014) or speech recognition (Amodei et al., 2016; Chan et al., 2016) triggered interest in applying such architectures to SLU, and after the introduction of the first approaches (Serdyuk et al., 2018; Haghani et al., 2018), the first public datasets for E2E SLU in English were created. The SNIPS benchmark (Coucke et al., 2018) and the Fluent Speech Commands (FSC) (Lugosch et al., 2019) were the first publicly released datasets, and due to the lack of datasets to compare with, FSC became the main corpus used as a benchmark for E2E SLU systems. However, the results obtained with this dataset were not representative of the actual performance of the technology, as it is a limited dataset in terms of number of sentences, recordings and participants, and is of low semantic complexity (McKenna et al., 2020). Therefore, efforts to create and share SLU datasets for E2E systems have continued. The next large corpus presented was the Spoken Language Understanding Resource Package (SLURP) (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 dataset, was presented 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 the TOP dataset for NLU (Gupta et al., 2018)).

In this paper we present the FalAI dataset [2,3], which, despite having fewer sentences than the SLURP and STOP datasets, is the largest publicly available dataset for E2E SLU to the best of our knowledge, in terms of hours, recordings and participants. FalAI is also the first dataset for E2E SLU in Galician, and the first such dataset for a low-resource language, which is an unprecedented milestone for the language. In addition, our dataset introduces novel splits for noisy audio, audio with hesitations, and audio where the structured information in the form of domain, intent, and slots is preserved, but the original sentence has been modified, providing a unique and comprehensive resource for evaluating E2E SLU systems under different real-world conditions.

2.1. Galician Context

Galician, which belongs to the Romance language family, is a co-official language, along with Spanish, in the autonomous region of Galicia, located in northwestern Spain, and has approximately 1.9 million speakers (I.G.E., sep 2019a). Moreover, Galician is a language that has linguistic variations depending on the geographical area and that 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 (Ramírez-Sánchez et al., 2022). As stated in the ELE (European Language Equality) report on Galician (Ramírez-Sánchez and García-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,

[2]The dataset can be found at: https://huggingface.co/datasets/GTM-UVigo/FalAI

[3]The dataset can also be accessed in the ELG: https://live.european-language-grid.eu/catalogue/corpus/21575

machine translation, text analysis or dialogue systems.

For the specific case of end-to-end spoken language understanding, there is no existing dataset of speech recordings for Galician, so FalAI is the first public dataset with 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 available datasets are the Common Voice dataset (Ardila et al., 2020), which in its version 15 for Galician has a total of 38 hours of validated speech, and the openSLR dataset (Kjartansson et al., 2020), which has about 10 hours of audio, so that in total there were 48 hours of high quality speech available for Galician.

3. FalAI Dataset

3.1. Dataset Design

The text corpus of the FalAI dataset was designed and created in collaboration with linguists to try to collect all the variants of the language, to correct and revise any errors that may have been 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 institution responsible for establishing the rules for the correct use of the language, the *Real Academia da Lingua Galega*.
- An effort has been made to balance gender and location 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 in order to increase diversity.
- In Galician there are situations in which two ways 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. We have decided to include both written versions in the corpus, since we noticed that participants did not pronounce correctly, as they read literally what was written.
- References to Galician culture and the way of life of the Galician people have been included in order to make the sentences attractive and to make people identify with them.

The FalAI dataset consists of a total of 3,500 sentences across 14 domains, with 64 intents and 62 slot types. The domains in the dataset include classic voice command phrases for smart home speakers, but also domains related to e-health, transport booking or questions about administrative processes. Table 1 shows a comparison between the text characteristics of the benchmark SLU datasets and FalAI dataset.

3.2. Data Collection

The FalAI data collection was conducted during a specific campaign in the first quarter of 2023. Citizens were invited to participate by recording themselves reading thirty sentences using a specially designed tool. The number of sentences per recording session was determined to strike a balance between a minimum number of recordings and participant motivation, typically taking 2-4 minutes per round. Participants were encouraged to use their natural Galician, acknowledging the language's rich variations beyond the standard form.

While the recording tool was primarily designed for mobile devices, it was accessible from any device with a microphone-equipped browser [4]. The design aimed for effectiveness and user-friendliness, ensuring accessibility for participants with varying technological capabilities. Figure 1 provides a visual representation of the main recording screen.

Figure 1: FalAI Recording Screen Example.

To participate in FalAI, users were required to provide information regarding their age, gender, place of origin, and accent. Additionally, they had to accept the data release document. This document was drawn up in collaboration with legal

[4]The FalAI recording tool can be accessed via: https://falai.balidea.com/

	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**
Vocab size	96	2,182	6,467	15,056	**2,957**

Table 1: Text corpora SLU dataset comparison.

experts and adheres to current recommendations and guidelines on data usage in virtual assistants (Piñeiro-Martín et al., 2023). There is no limit to the number of participations, and each unique contribution counts as an additional participation, as we do not retain personal data or associated identifiers of participants.

The data collection campaign has been an unprecedented success for the language, surpassing all expectations. It has resulted in an astounding 6 times more hours of audio data compared to the main existing speech datasets in Galician. This campaign has achieved remarkable representation, with the participation of 99% of the municipalities in Galicia, with recordings from 311 of the 313 Galician municipalities. The dataset includes more than 25,000 recordings per province and also captures the rich diversity of Galician accents, with more than 25,000 recordings for each of the main accent variations. In particular, the campaign succeeded in involving a significant number of participants over the age of 60, a demographic typically underrepresented in such initiatives, with more than 15,000 recordings from this age group.

Table 2 summarises the main results of the FalAI data collection campaign:

Number of hours	250
Number of recordings	260,000+
Number of participants	10,000+
Municipalities participating	99%
Female / Male ratio	60% - 40%
Hours from participants aged 60+	18.3

Table 2: Main results of the FalAI data collection campaign.

3.3. Dataset Validation

The validation process for the FalAI dataset was carefully designed due to the unique challenges posed by Galician, including the sensitivity of meaning to single character changes and the complexity of sentences containing numbers, municipality names or acronyms. Unlike English ASRs, Galician ASRs do not have comparable confidence scores. As a result, we opted for a semi-automatic validation strategy, as opposed to the fully automatic approach used in SLURP and the semi-automatic approach with a 50% character error rate (CER) automatic validation threshold in STOP. Given the complexity of Galician, it is not feasible to replicate such a low threshold.

The validation strategy of the FalAI dataset included five different phases. Throughout the validation process, all recordings were amplitude normalised and initial and final silences were identified and removed:

1. **Manual validation**: In the first phase, 12,750 recordings (approximately 5% of the dataset) underwent manual validation, creating a representative corpus with fully supervised validation.

2. **ASR fine-tuning:** In the second phase, the XLS-R model (Babu et al., 2022) was fine-tuned using Galician speech data from Common Voice (Ardila et al., 2020) and OpenSLR (Kjartansson et al., 2020). About 30% of the dataset (75,000 recordings) with 0% word error rate (WER) were automatically validated.

3. **Language model boost**: In the third phase, a 4-gram model trained on the dataset text corpus was used as the language model to boost recognition. In this phase, an additional 10% of the original dataset was validated with zero WER.

4. **ASR model refinement**: In the fourth phase, the XLS-R model was further fine-tuned using the recordings validated in the first phase. This new ASR model, reinforced by the 4-gram model, validated an additional 30% of the original dataset, with 75% of the dataset validated by the end of this phase.

5. **Manual validation (final phase)**: The final phase manually validated recordings that hadn't been validated in previous phases.

All automatically validated audios have been labelled "*validated*", while in the manual validation we have included additional labels to create splits that provide a unique resource for evaluating E2E SLU systems in new scenarios.

3.3.1. FalAI Splits

The FalAI dataset introduces novel splits generated during manual validation using additional labels. The labels used during manual validation are:

- **Validated**: Utterance that exactly matches the reference sentence.
- **Changed**: Utterance in which some word(s) have been changed or pronounced differently, but the semantic information in terms of intent and slots is retained.
- **More words**: Utterance in which words are added, but the structured information is retained.
- **Fewer words**: Utterance in which words are omitted but the structured information is retained.
- **Hesitation**: Utterance in which there is a hesitation in pronunciation, whether or not the reference sentence has changed, but in which the semantic information is retained.
- **Noisy**: Noisy recording, background noise or audio problems.

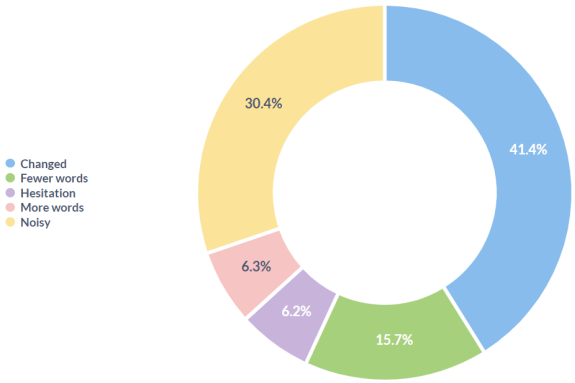


Figure 2: Distribution of additional splits in FalAI.

In the first released version of the validated FalAI dataset, these splits account for the 10% of the total dataset, representing more than 24 hours of audio in total. These splits will allow the evaluation of the E2E models in challenging scenarios, testing their performance at the boundaries and assessing their ability to adapt to the different linguistic variations present in Galician. This includes testing their robustness in noisy environments, their ability to handle changes in the original message and their performance in more realistic scenarios with the presence of hesitations. Figure 2 shows the distribution within these additional splits.

4. Dataset Complexity

In the domain of Spoken Language Understanding (SLU), assessing the complexity of datasets is a critical aspect of designing effective evaluation benchmarks. However, relying solely on near-perfect results obtained with limited datasets such as FSC or SNIPS can be misleading, as these datasets often lack the lexical and semantic richness, diversity of vocalisations, domain coverage and diverse semantic contexts encountered in real-world scenarios (Bastianelli et al., 2020). To bridge this gap and provide a comprehensive assessment of dataset complexity, this section aims to present key metrics for evaluating the textual complexity of corpora. Furthermore, we introduce novel metrics based on our practical experience to further enhance the understanding of dataset complexity.

4.1. 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 sentences in the dataset over its constituent n-grams, $\mathcal{N}$. It is calculated using the equation:

$$H = -\sum_{x \in \mathcal{N}^*} p(x) log_2 p(x) \quad (1)$$

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 greater randomness and variety in the utterance patterns, indicating greater lexical complexity. Table 3 shows a comparison between the entropy for unigrams, bigrams, trigrams and the average entropy between the main benchmark datasets in the literature: the Fluent Speech Commands (FSC), SNIPS, SLURP and STOP datasets and the FalAI dataset.

Entropy	FSC	SNIPS	SLURP	STOP	**FalAI**
1-gram	5.5	6.2	8.8	9.2	**9.3**
2-gram	7.2	9.1	13.1	13.6	**12.5**
3-gram	7.9	10.9	14.7	15.9	**13.4**
average	6.9	8.7	12.2	12.9	**11.7**

Table 3: Comparison of entropies between the main SLU datasets and the FalAI dataset.

4.2. Semantic Analysis: Semantic Textual Similarity

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, summarisation, question answering or semantic search. In this section, we propose to compute the STS as a measure of the semantic complexity of the designed dataset.

To compute the STS, the first step is to obtain the vector representations (embeddings) of the sentences in the corpus. For this purpose, the Language-agnostic BERT Sentence Encoder (LaBSE) (Feng et al., 2022), based on BERT (Devlin et al., 2019) and trained for more than 100 languages (including Galician, Spanish and English), was used. After that, an L2 normalisation (Feng et al., 2022) is performed to avoid that parameters of vectors with different ranks and high variance influence more than those that are not normalised, and then the similarity is calculated as the product between the two tensors.

The STS heatmap is a visual representation that provides valuable insights into the degree of semantic equivalence between different elements within a dataset. In this context, the STS heatmap serves as a powerful tool for assessing the linguistic and semantic richness of the dataset, helping us to understand how diverse or homogeneous the dataset is at different levels of granularity.

Figure 3 shows a heatmap of the STS at the domain level, with 1 being the highest similarity and 0 the lowest similarity. Although there is some relationship between some of the domains, the overall similarity of the map shows that the domains are diverse and broad, indicating the semantic richness of the dataset.

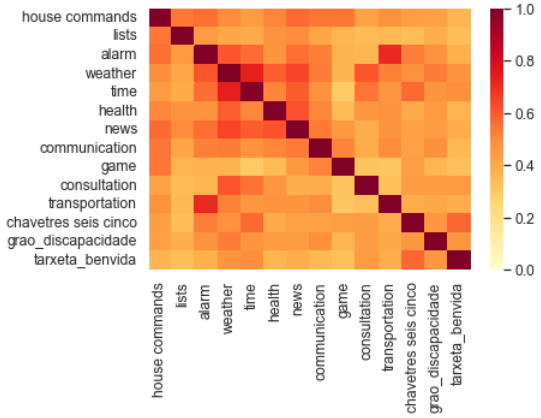
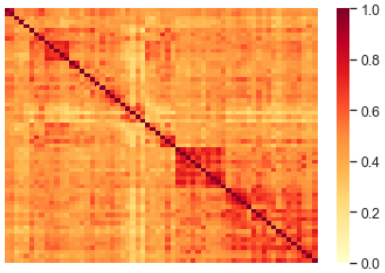


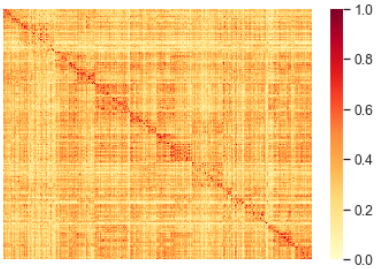
Figure 3: STS between domains in FalAI calculated with LaBSE.

The STS at the level of intents (Figure 4(a)) in the dataset shows that the average value is higher, which makes sense since the way of requesting information or making requests tends to be similar regardless of what is being requested. This higher STS may also indicate a higher complexity in classifying intents (the more similar they are, the more difficult it is to classify them).

Finally, Figure 4(b) shows the STS between sentences in the dataset. This heatmap illustrates the varying degrees of similarity between different sentences. In particular, there are regions with higher STS values, typically associated with individual sentences or small groups of sentences. As expected, the average STS decreases as the number of sentences compared increases. This diversity reflects the richness of the dataset, as it encompasses a wide range of linguistic variations and potential semantic nuances.



(a) STS between intents.



(b) STS between sentences.

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

The observed STS patterns across domains, intents and sentences provide valuable insights into the multifaceted nature of the dataset. While the lower STS values between certain domains indicate diversity and distinctiveness among them, the elevated STS at the intent level suggests intricate semantic relationships between different intents. At the same time, the lower average STS at the sentence level indicates a dataset that contains a wide range of linguistic variation and potential semantic nuance. These combined findings highlight the complexity of the dataset, making it a versatile resource for evaluating and improving spoken language understanding systems in different real-world scenarios.

	FSC	SNIPS	SLURP	STOP	**FalAI**
Speakers	97	67	177	885	**10,000+**[5]
Audio files	30,043	5,886	72,277	236,477	**260,000+**
Duration [hrs]	19	5.5	58	218	**250**

Table 4: Comparison of speech data between datasets.

5. Results and Analysis

FalAI represents a major milestone in the field of spoken language understanding. Not only is it the largest public dataset for SLU in any language, but it is also the first dataset for SLU in Galician, which represents a significant step forward for the language in terms of linguistic resources, and the first dataset for SLU obtained in a low-resource scenario. Table 4 shows the comparison of speech data between the main datasets in the literature and the FalAI dataset.

The release of FalAI means going from a few tens of hours of voices to hundreds of hours, with thousands of speakers from all regions of Galicia. The metadata associated with the recordings will also be of great use for future research, as it is a dataset in which the variety of accent is recorded and efforts have been made to ensure that participants use their variety of Galician. In addition, the dataset has an important representation of the voices of people over the age of 60, a population group that is typically underrepresented in speech datasets, and which is even more important for Galician, since in Galicia these people are the most likely to use variation of the Galician language other than the normative one.

FalAI is also the first dataset for SLU that presents novel splits related to noisy audio, audio with hesitations, or audio where the reference sentence has been modified but the semantic information is preserved in the form of intents and slots (about 24 hours of audio in this first release). This last split is particularly interesting for testing the generalisability of future E2E SLU systems, since they are recordings in which the same structured information is expressed as in the reference sentence, but the way in which it is expressed is changed, either by using more or fewer words, by using a different variety of language, or simply by changing the way in which the information is requested.

Moreover, in our comprehensive assessment of the complexity of the FalAI dataset, we have examined key metrics for evaluating both lexical and semantic complexity. This analysis provides essential context for understanding the richness and diversity of the dataset, which in turn plays a crucial role in evaluating the performance of SLU systems on real-world data.

Finally, the process of dataset design, collection and validation in a low-resource scenario also provides important lessons for the creation of datasets in similar contexts through citizen collaboration. Some of the lessons learned can be summarised as follows:

- **Transparent communication and clear basis**: Clear communication of the objectives of the data collection campaign, as well as the terms and conditions of data transfer through the data release document, has helped to build citizens' trust in the project.
- **Public-private collaboration**: The project was carried out thanks to a public-private collaboration between a technology company and a research centre linked to a university. We believe that this collaboration has been key to the success of the project, leveraging the impact, image and knowledge of the public side and the implementation capacity and experience of the private side.
- **The reason**: The campaign was promoted as an opportunity to ensure that Galician was also present in the digital world. Getting people to feel involved in this proposal and to understand that it was necessary for the language was the main motivating element of the campaign.
- **Linguistic content of the dataset**: We have also seen how the linguistic content of the dataset has also helped to extend the reach of the campaign. References to Galician culture (writers, actors or musicians) as well as Galician habits, traditions and regions have helped to popularise the campaign and encourage participation, i.e. people have participated because they felt reflected and found it useful, illustrating the importance of developing people-centred applications.
- **The recording tool**: We have taken great care in the design and development of the recording tool to make it simple and intuitive, regardless of the technology profile. We believe that having a stable, multi-device, intuitive, attractive and well-maintained tool was another key factor in the project.

[5]The exact total number of participants is an estimate due to the lack of identifying information for each participant.

- **Social networks**: The impact of social media has undoubtedly been the most effective way to publicise FalAI's campaigns. Collaborating with local influencers and celebrities has helped to increase engagement with the campaign.

6. Conclusions

In this paper we present FalAI, the largest SLU dataset in terms of hours, recordings and participants across all languages. FalAI is also the first SLU dataset for Galician and significantly expands the linguistic resources available for what has traditionally been considered a low-resource language, increasing the resources by more than a factor of six. Furthermore, FalAI introduces novel dataset splits not previously seen in the SLU literature, providing a unique resource for evaluating end-to-end SLU systems in diverse real-world scenarios. This versatility sets the stage for innovative research in the field.

In addition to its size and diversity, this work has been concerned with evaluating the textual complexity of FalAI. We have proposed measures of both lexical and semantic complexity, providing valuable insights into the richness of the dataset and its potential for extensive SLU research.

The process of designing, collecting and validating FalAI has resulted not only in a remarkable dataset, but also in valuable lessons for the research community. These lessons highlight the importance of transparent communication, public-private collaboration, a good recording tool, and making people feel involved and represented.

In conclusion, FalAI represents a significant advancement in the field of E2E SLU, not only for Galician but across all languages. With its extensive scale, linguistic diversity, and innovative evaluation possibilities, it offers valuable avenues for future research and the creation of datasets in similar scenarios. As a part of our future work, we plan to test the dataset by comparing traditional SLU architectures with E2E architectures, particularly focusing on the new splits introduced in FalAI.

7. Acknowledgements

The FalAI project was funded by the European Language Equality 2 project (ELE 2), which has received funding from the European Union under the grant agreement no. LC-01884166 – 101075356.

We would like to thank the Galician Innovation Agency (GAIN) and the Consellería de Cultura, Educación, Formación profesional e Universidades of the Xunta de Galicia for funding the industrial PhD through the program: Doutoramento Industrial.

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

8. Bibliographical References

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pages 173–182. PMLR.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Proc. Interspeech 2022*, pages 2278–2282.

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

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. Slurp: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262.

Valentina Bellomaria, Giuseppe Castellucci, Andrea Favalli, and Raniero Romagnoli. 2019. Almawave-slu: A new dataset for slu in italian. *arXiv preprint arXiv:1907.07526*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. Semeval-2017 task 1: Semantic textual similarity multilingual

and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14.

William Chan, Navdeep Jaitly, Quoc Le, and Oriol Vinyals. 2016. Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Regina Gubareva and Rui Pedro Lopes. 2020. Virtual assistants for learning: A systematic literature review. *CSEDU (1)*, pages 97–103.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2787–2792.

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

Charles T Hemphill, John J Godfrey, and George R Doddington. 1990. 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*.

I.G.E. sep 2019a. Enquisa estrutural a fogares. coñecemento e uso do galego. resumo de resultado 27/09/2019.

Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Oddur Kjartansson, Alexander Gutkin, Alena Butryna, Isin Demirsahin, and Clara Rivera. 2020. Open-Source High Quality Speech Datasets for Basque, Catalan and Galician. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 21–27, Marseille, France. European Language Resources association (ELRA).

Loren Lugosch, Mirco Ravanelli, Patrick Ignoto, Vikrant Singh Tomar, and Yoshua Bengio. 2019. Speech Model Pre-Training for End-to-End Spoken Language Understanding. In *Proc. Interspeech 2019*, pages 814–818.

Joseph P. McKenna, Samridhi Choudhary, Michael Saxon, Grant P. Strimel, and Athanasios Mouchtaris. 2020. Semantic Complexity in End-to-End Spoken Language Understanding. In *Proc. Interspeech 2020*, pages 4273–4277.

Andrés Piñeiro-Martín, Carmen García-Mateo, Laura Docío-Fernández, and María del Carmen López-Pérez. 2023. Ethical challenges in the development of virtual assistants powered by large language models. *Electronics*, 12(14):3170.

Shangeth Rajaa, Swaraj Dalmia, and Kumarmanas Nethil. 2022. Skit-s2i: An indian accented speech to intent dataset. *arXiv preprint arXiv:2212.13015*.

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

José Manuel Ramírez-Sánchez and Carmen García-Mateo. 2022. Deliverable D1.15 Report on the Galician Language.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27.

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

Francesco Vona, Emanuele Torelli, Eleonora Beccaluva, and Franca Garzotto. 2020. Exploring the potential of speech-based virtual assistants in mixed reality applications for people with cognitive disabilities. In *Proceedings of the International Conference on Advanced Visual Interfaces*, pages 1–9.

Su Zhu, Zijian Zhao, Tiejun Zhao, Chengqing Zong, and Kai Yu. 2019. Catslu: The 1st chinese audio-textual spoken language understanding challenge. In *2019 International Conference on Multimodal Interaction*, pages 521–525.