BALiDEA

atlanTTic
Universidade Vigo

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, Spain
[2]Balidea Consulting & Programming S.L., Santiago de Compostela, Spain

LREC-COLING 2024

Outline

01 MOTIVATION

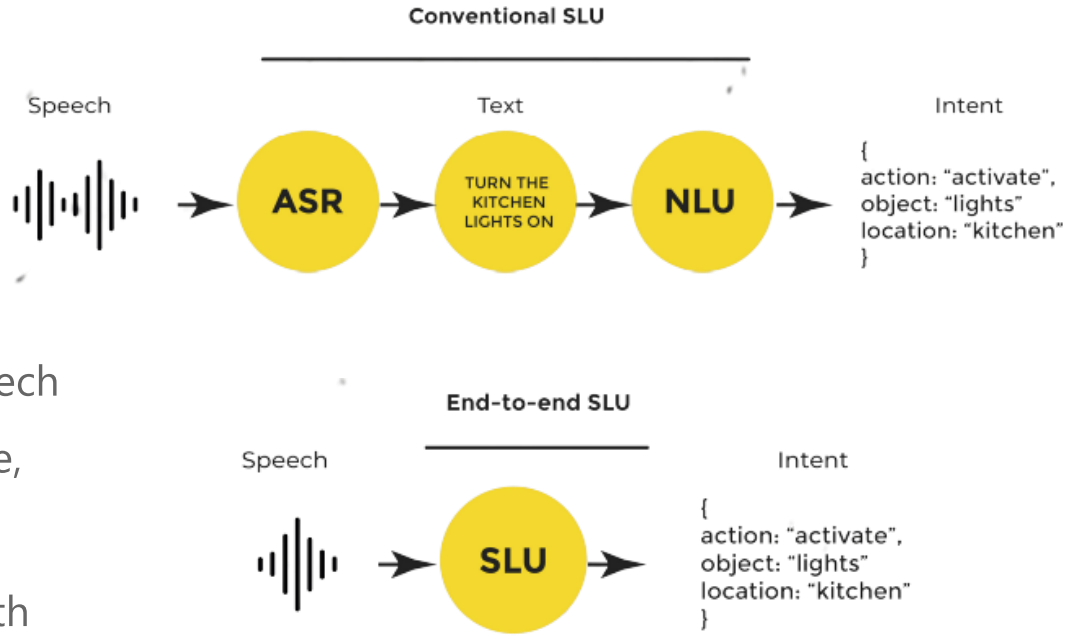
02 TEXTUAL DATASET DESIGN

03 DATASET COLLECTION AND VALIDATION

04 RESULTS

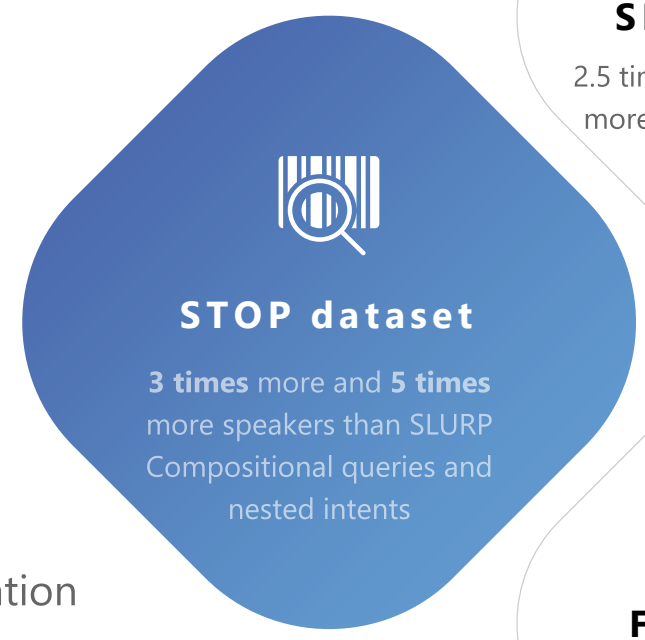
Motivation

- **Spoken Language Understanding** (SLU) **->** structured information from speech signals.
- **E2E architectures ->** information extracted directly from speech signal **->** Prevent cascading errors, optimized in a single stage, leverage prosody, etc.
- **Problem ->** lack of large and challenging speech datasets with structured information or semantic parsing labels **->** **low complexity** and **limited domain**.



Related work

- **ATIS** (1990s) **->** first speech dataset with annotated structured information.
- **SNIPS** (2018) & **FSC** (2019) **->** benchmark dataset in SOTA E2E SLU.
- **SLURP** (2020) and **STOP** (Meta, 2023) **->** next generation of public datasets.
- Challenge even grater for **languages other than English ->** public datasets for Mandarin Chinese, Indian or Italian.



SLURP dataset

2.5 times more audio than FSC, more domains, greater lexical richness

FalAI dataset

What about Galician?

Galician context

- Co-official language in Galician.
- 1.9 million speakers.
- Linguistic variations, bilingualism and code-switching.
- Low-resource language.
- No datasets available for E2E SLU, and scarce speech resources (48 hours of labelled data).

BALiDEA

FalAI dataset

- **Largest publicly available dataset for SLU** in terms of hours, recordings and participants, **in any language**.
- **First SLU dataset** and **largest speech dataset** for **Galician**.
- 14 domains, 62 intents, 64 slots types with +1,8000 different values.
- Novel 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.

Textual dataset **design**

- **3,500 sentences** designed in collaboration with linguists.
- Gender and locations references balanced.
- Include references to the Galician culture.
- Typical virtual assistant domains such as **house commands**, **weather**, **alarms**, **lists**, but also domains such as **health** or **e-government.**

	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.

Textual dataset **complexity**

Lexical Analysis: n-gram Entropy

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

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**

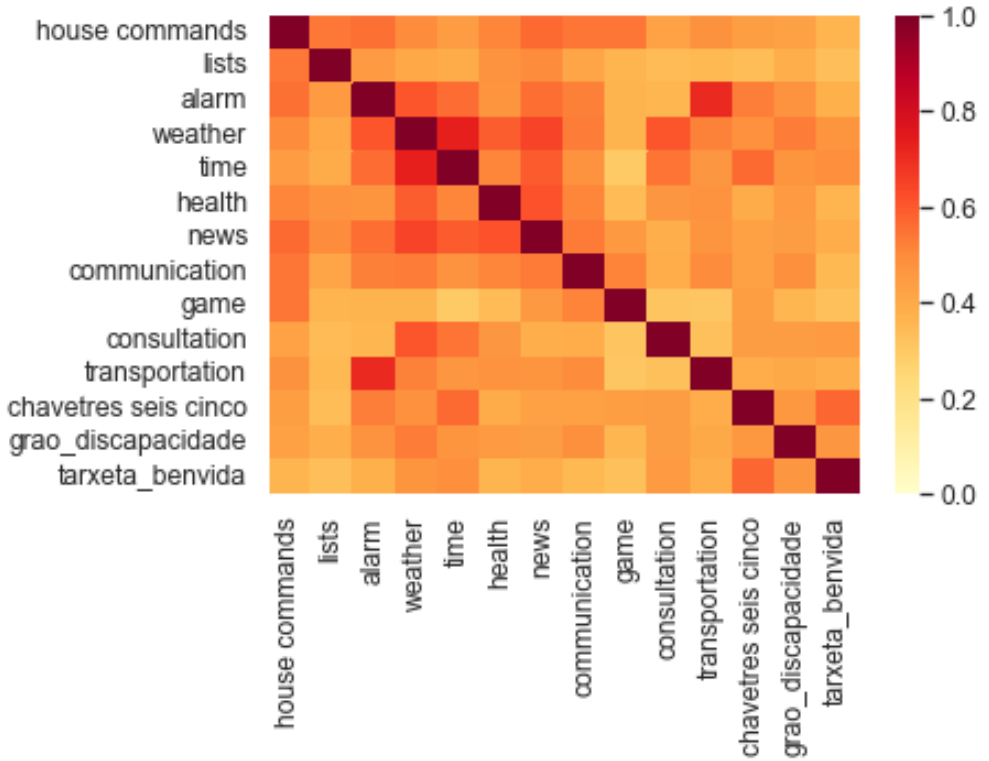
Table 3: Comparison of entropies between the main SLU datasets and the FalAI dataset.

Textual dataset **complexity**

Semantic Analysis: Semantic Textual Similarity

- Degree to which two sentences are semantically equivalent to each other.
- Calculated used **Language-agnostic BERT Sentence Encoder (LaBSE)**

STS between domains

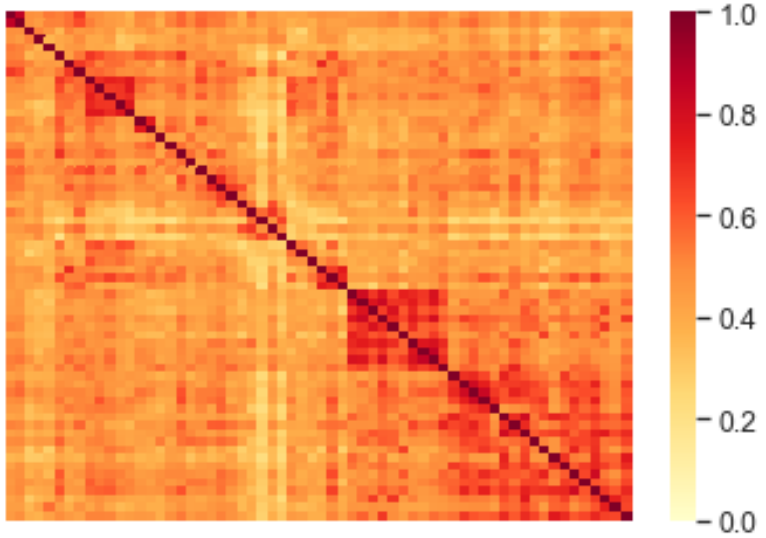


Textual dataset **complexity**

Semantic Analysis: Semantic Textual Similarity

- Degree to which two sentences are semantically equivalent to each other.
- Calculated used **Language-agnostic BERT Sentence Encoder (LaBSE)**

STS between intents

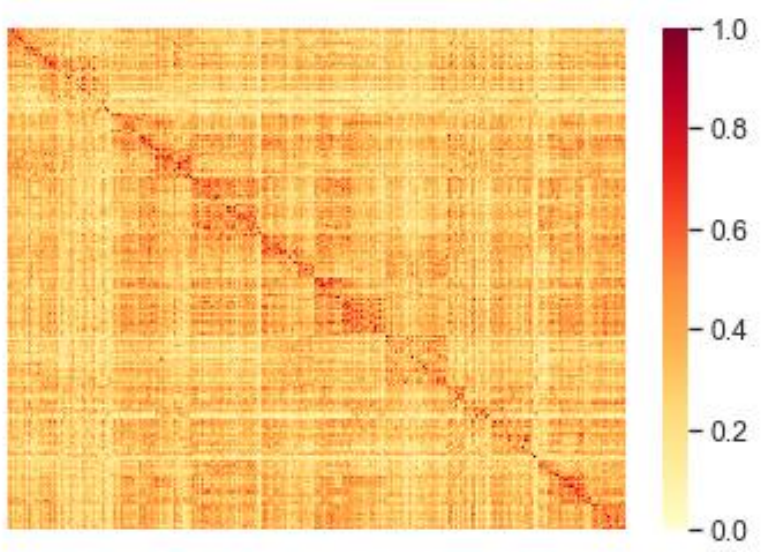


Textual dataset **complexity**

Semantic Analysis: Semantic Textual Similarity

- Degree to which two sentences are semantically equivalent to each other.
- Calculated used **Language-agnostic BERT Sentence Encoder (LaBSE)**

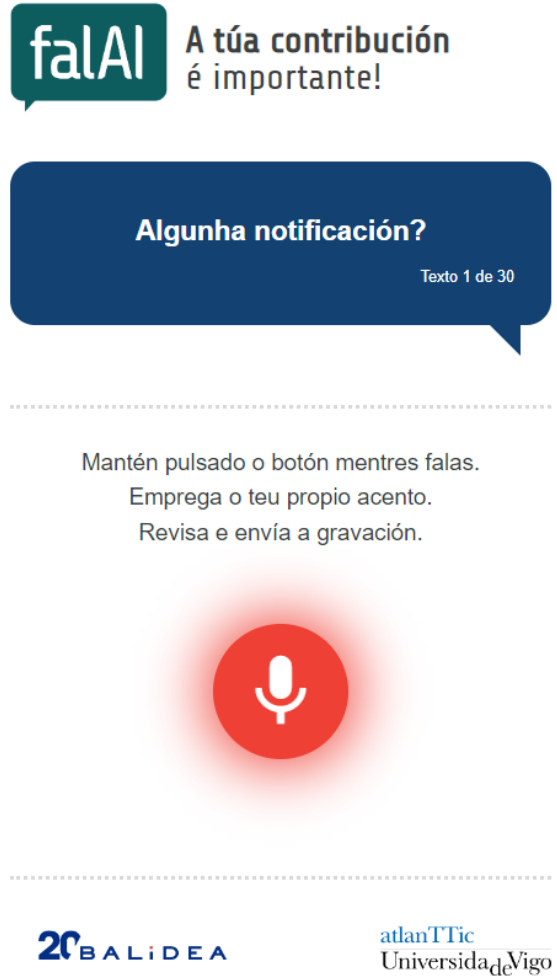
STS between sentences



Dataset **collection**

- Campaign in the first quarter of 2023.
- Citizens were invited to participate by recording themselves reading 30 sentences.
- We designed a tool for it, accessible from any device with a browser:

https://falai.balidea.com/



Collection **results**

- Unprecedented success for the language.
- **6 times more hours** of audio than the main Galician speech datasets.
- Participation of **99% of the municipalities**.
- More than **15,000 recordings** from participants over the age of **60**.

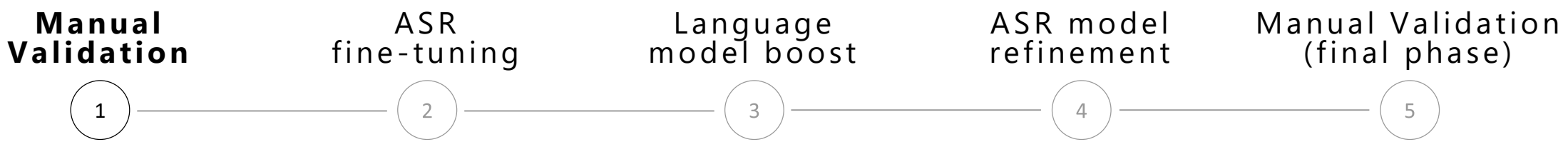
Number of hours	250
Number of recordings	260,000+
Number of participants	10,000+
Municipalities participating	99%
Female / Male ratio	60% - 40%
Hours from participants aged 60+	18.3

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

BALiDEA

atlanTTic
Universidade Vigo

Dataset **validation**



12,750 recordings **-> 5% of the dataset**

BALiDEA

atlanTTic
Universidade Vigo

Dataset **validation**

Manual Validation (1) — **ASR fine-tuning** (2) — Language model boost (3) — ASR model refinement (4) — Manual Validation (final phase) (5)

XLS-R model was fine-tuned using Common Voice

and OpenSLR.

75,000 recordings **-> 30% of the dataset**

with 0% WER automatically validated.

BALiDEA

atlanTTic
Universidade Vigo

Dataset **validation**

Manual Validation (1) — ASR fine-tuning (2) — **Language model boost** (3) — ASR model refinement (4) — Manual Validation (final phase) (5)

4-gram model trained with FalAI textual dataset

10% of the original dataset was validated with

0% WER.

BALIDEA

atlanTTic
Universidade Vigo

Dataset **validation**

Manual Validation (1) — ASR fine-tuning (2) — Language model boost (3) — **ASR model refinement** (4) — Manual Validation (final phase) (5)

XLS-R further fine-tuned using recordings validated in the first phase

30% of the dataset was validated with

0% WER **->** 75% of the original dataset was validated at the end of this phase

BALIDEA

atlanTTic
Universidade Vigo

Dataset **validation**

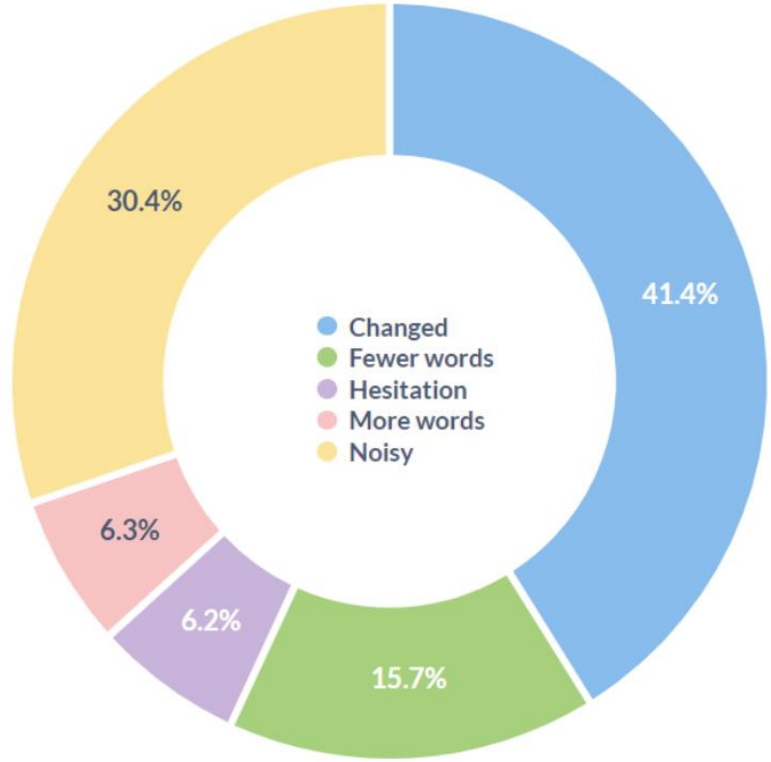
Manual Validation (1) — ASR fine-tuning (2) — Language model boost (3) — ASR model refinement (4) — **Manual Validation (final phase)** (5)

25% of the dataset was manually validated.

Splits were created

FalAI **Splits**

- **Validated**: exactly matches the reference sentence.
- **Changed**: some word(s) have been changed or pronounced differently, but information is retained.
- **More words**: words are added but information is retained.
- **Fewer words**: words are omitted but information is retained.
- **Hesitation**: hesitation in pronunciation.
- **Noisy**: noisy recording, background noise or audio problems.



Results

- Galician -> from few tens of hours to **hundreds of hours, with thousands of speakers.**
- **Novel splits** not previously seen in SLU literature (**noisy, hesitation** or **changed** splits) -> test E2E SLU systems.
- **Valuable metadata ->** accent, gender, age, location.

	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.

Conclusions

- **Largest publicly available dataset for SLU** in terms of hours, recordings and participants, **in any language.**
- **Lessons learned** through the design, collection and validation.
- **Lexical and semantic** complexity measures.
- Potential for extensive **SLU research**.

Hugging Face

https://huggingface.co/datasets/GTM-UVigo/FalAI

```
from datasets import load_dataset

falai = load_dataset("GTM-UVigo/FalAI", split="validated")
```

BALiDEA

atlanTTic
Universidade Vigo

Thank you

Andrés Piñeiro-Martín
andres.pinerio@balidea.com

LREC-COLING 2024