Developing Speech Corpora for Low-Resource Languages

Prof. Gražina Korvel
grazina.korvel@mif.vu.lt

The data diversity in machine learning

Developing diverse, well-annotated speech corpora is essential for training modern machine learning models.

✓However, most progress in AI has focused on a few ***high-resource languages***.

The goal of the presentation

This presentation discusses the principles and methodologies of creating large-scale speech corpora, focusing on the *Lithuanian language* as a case study.

I am currently involved in three projects that focus on developing and annotating corpora

Project "Creation of the Great Lithuanian Speech Corpus (LIEPA-3)",
Project Nr. 02-023-K-0001, 12.03.2024 - 30.04.2026.

✓***10 000 hours of annotated Lithuanian speech for speech recognition***

Project "Creation of a Lithuanian Speech Corpus (for speech synthesis purposes to generate neural voices)“, No. (1.57) 15600 INS 70, 16.10.2024 - 30.04.2026.

✓***200 hours of annotated Lithuanian speech for synthesis + 300 hours of annotated Lithuanian emotional speech***

Project "Research on propaganda and disinformation: automatic recognition by machine learning, impact and societal resilience", No. S-VIS-23-8, 2023-09-01-2026-06-30

✓***1 000 social network news articles, annotated for propaganda techniques and narratives***

Project LIEPA

“Services managed by Lithuanian language”

LIEPA *2013.02 – 2015.08*

LIEPA2 *2017.12 – 2020.12*

Key Outcomes:

- **Lithuanian Speech Recognition System**: A functional speech recognition system for Lithuanian.
- **Lithuanian Text-to-Speech System**: TTS system for converting Lithuanian text into speech.
- **Speech Corpora**: Annotated data resources for Lithuanian speech research and development.

100 hours of speech data ***1000 hours*** of speech data were collected. The corpus was designed to cover a wide range of phonetic variations in Lithuanian language, including all vowels, consonants, and diphthongs.

Project LIEPA-3

„ Creation of the Great Lithuanian Language Corpus“

12 March 2024 to 30 April 2026

The project aims to develop an annotated Lithuanian language corpus of at least ***10000 hours*** for speech recognition, AI, and innovative technologies development

Vilniaus universitetas

VYTAUTO DIDŽIOJO UNIVERSITETAS
MCMXXII

LIETUVIŲ KALBOS INSTITUTAS

Recording Specifications

Recordings must be stored at: ***44 kHz*** sampling rate and ***16-bit*** mono format.

Originality and uniqueness

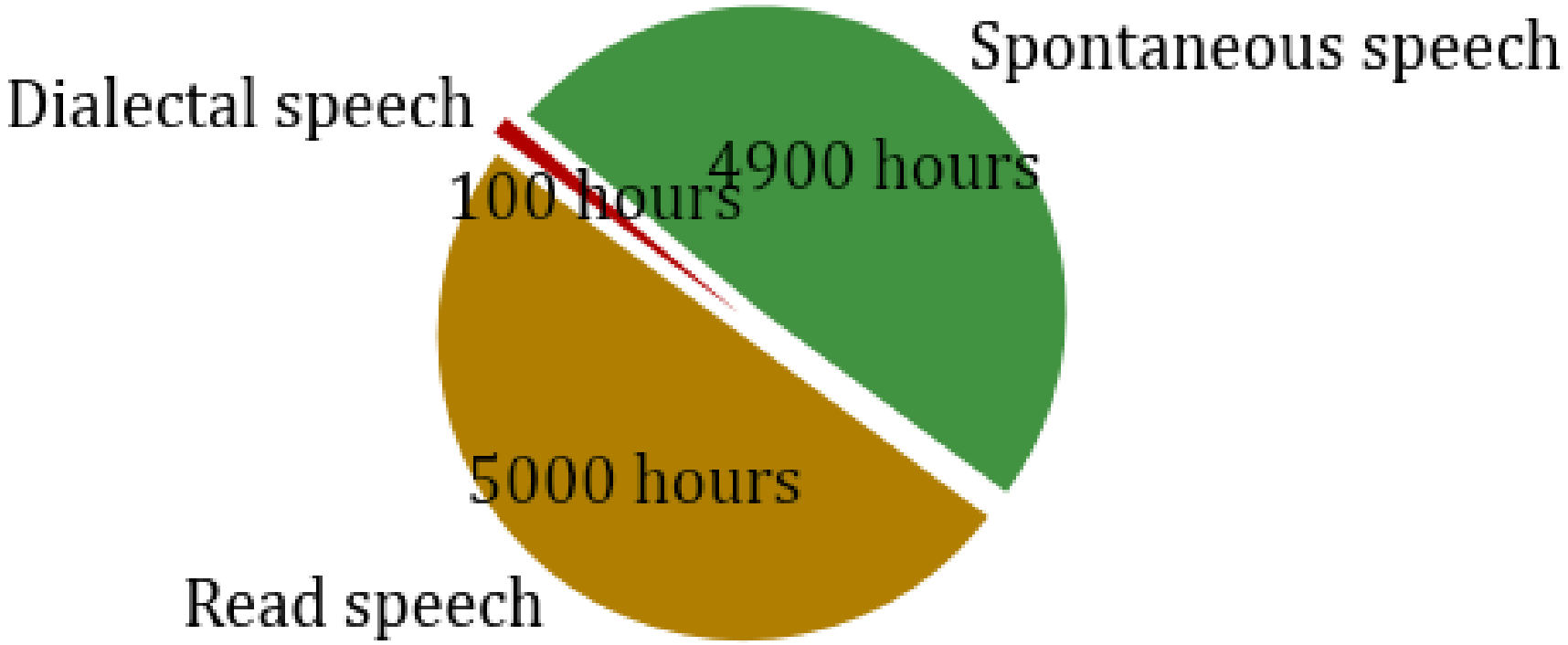
- ✓The corpus must be a new linguistic resource.
- ✓It cannot reuse or embed any previously created Lithuanian corpora.
- ✓No duplicate recordings are allowed.

Dialectal speech

- ✓The corpus must include dialectal speech samples representing the following dialects:
  - *Žemaitian*
  - *Dzūkian*
  - *Suvalkian*
  - *Aukštaitian*

Minimum: 100 hours in total.

Stylistic diversity



Recording Methodology in the LIEPA-3 Project

Read speech recordings come from two sources:

- ✓direct in-person sessions in Lithuanian cities, and
- ✓the *https://kurkgarsyna.lt/* online platform, where volunteers record speech remotely.

Spontaneous speech is collected from existing radio broadcasts under official collaboration agreements.

Dialectal speech samples are collected through in-person recordings.

Key features of corpus

- ✓The corpus consists of Lithuanian speech recordings (***non-Lithuanian language content limited to 0.1%***).
- ✓ ***High-quality, noise-varied*** recordings.

Non-standard language

The corpus must include speakers using ***non-standard Lithuanian***, such as slang, or obscene language.

✓Minimum: *100 hours of such recordings*.

Corpus composition must reflect

- ✓Gender, age, and dialect region of speakers.
- ✓***Phonetic diversity***: balanced coverage of Lithuanian sounds and combinations.
- ✓***Topical diversity***: speech across various themes.

Age groups used in our corpus

Gender	Age groups
F (Female) M (Male)	1: 0-12 years old 2: 13-17 years old 3: 18-60 years old 4: 60+ years old

Recording conditions

Must include recordings captured:

- ✓With ***professional*** and ***consumer*** recording ***equipment***.
- ✓In different ***acoustic environments***.
- ✓Under various ***noise conditions***.

Annotation requirements

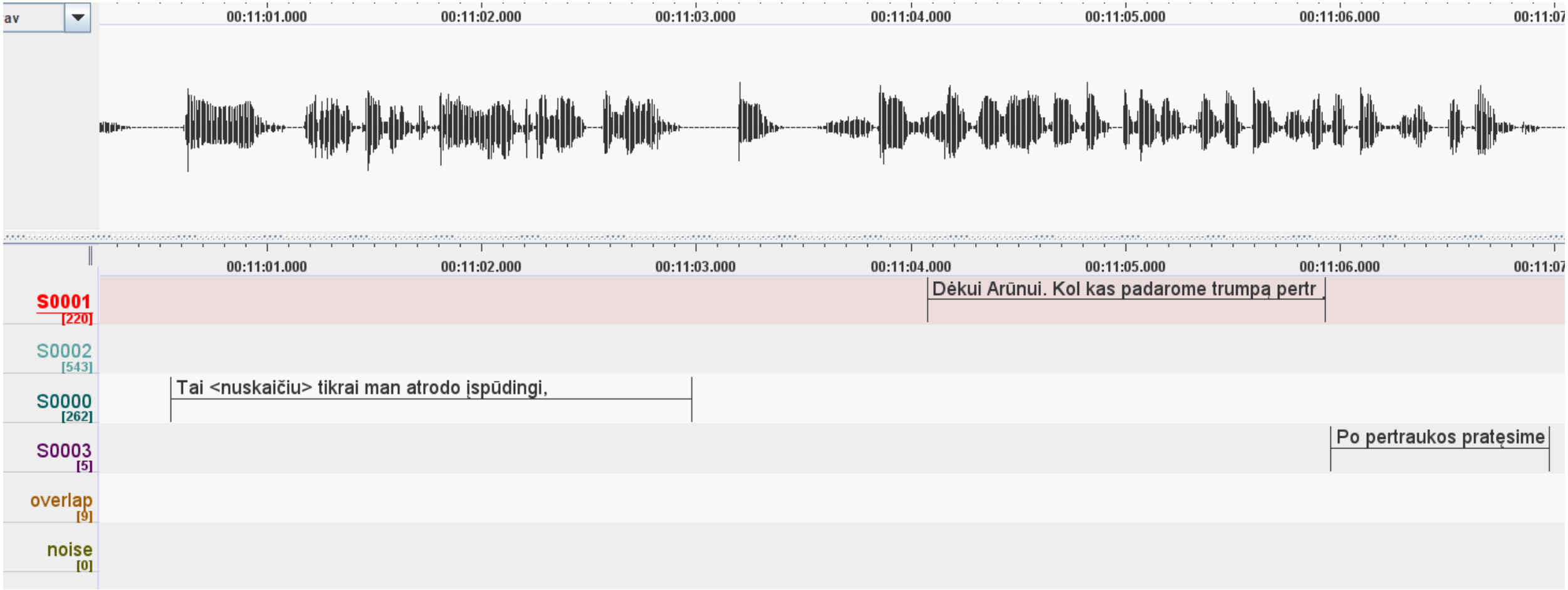
All recordings in the corpus must be annotated at

- ✓ ***The utterance level***:
  - The text of the spoken phrase.
  - The start and end timestamps of that phrase.

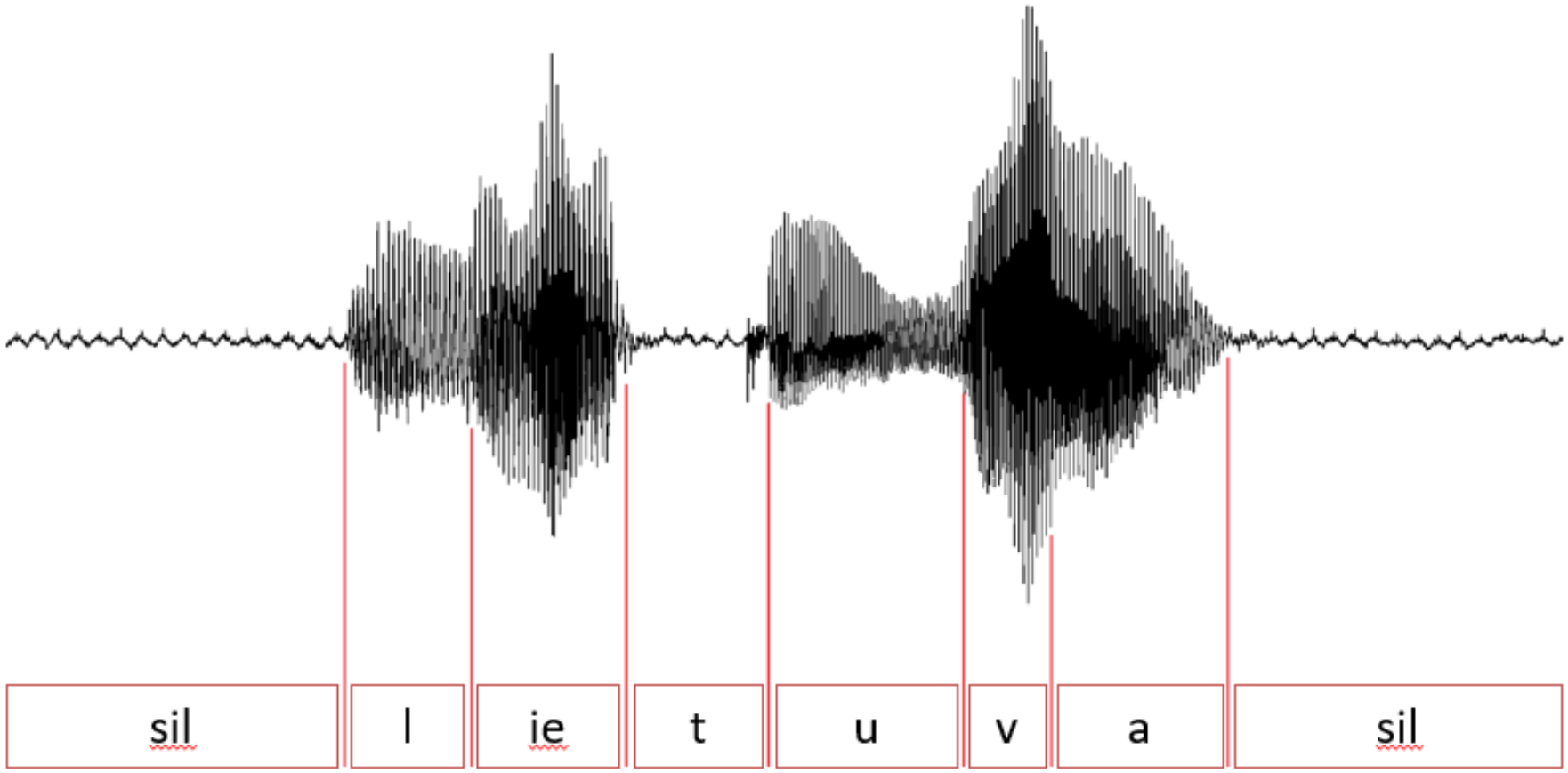
At least 500 hours of recordings must be annotated at

- ✓ ***The lexical unit level*** (individual words).
- ✓ ***The phoneme level*** (individual speech sounds).

The utterance level annotation



The phoneme level annotation



Corpus validation requirements

Purpose of Validation:

- ✓To ensure that the ***annotations accurately match the audio recordings***.
- ✓To demonstrate that the corpus ***is reliable and suitable for developing speech recognition technologies***

Annotation validation process

A formal annotation validation procedure must be performed.

- ✓The level of mismatch between audio and annotation must not exceed: ***0.1% at the utterance level***, and ***0.05% at the lexical level***.

Demonstration of usability

Use 1% of the total corpus (minimum 100 hours) to build a demonstration speech recognition system.

✓The system's ***Word Error Rate*** (WER) must be ≤ 20%.

$$\mathrm{WER} = \frac{S + D + I}{N} \times 100\%$$

S – number of substitutions (incorrect words),
D – number of deletions (missing words),
I – number of insertions (extra words not in the reference)
N – total number of words in the reference (ground truth)

Open Access and Distribution

The corpus must be openly accessible under a Creative Commons license.

Must be published on at least two ***open-access platforms***, such as:

- Hugging Face
- CLARIN
- Lithuanian Open Data Portal

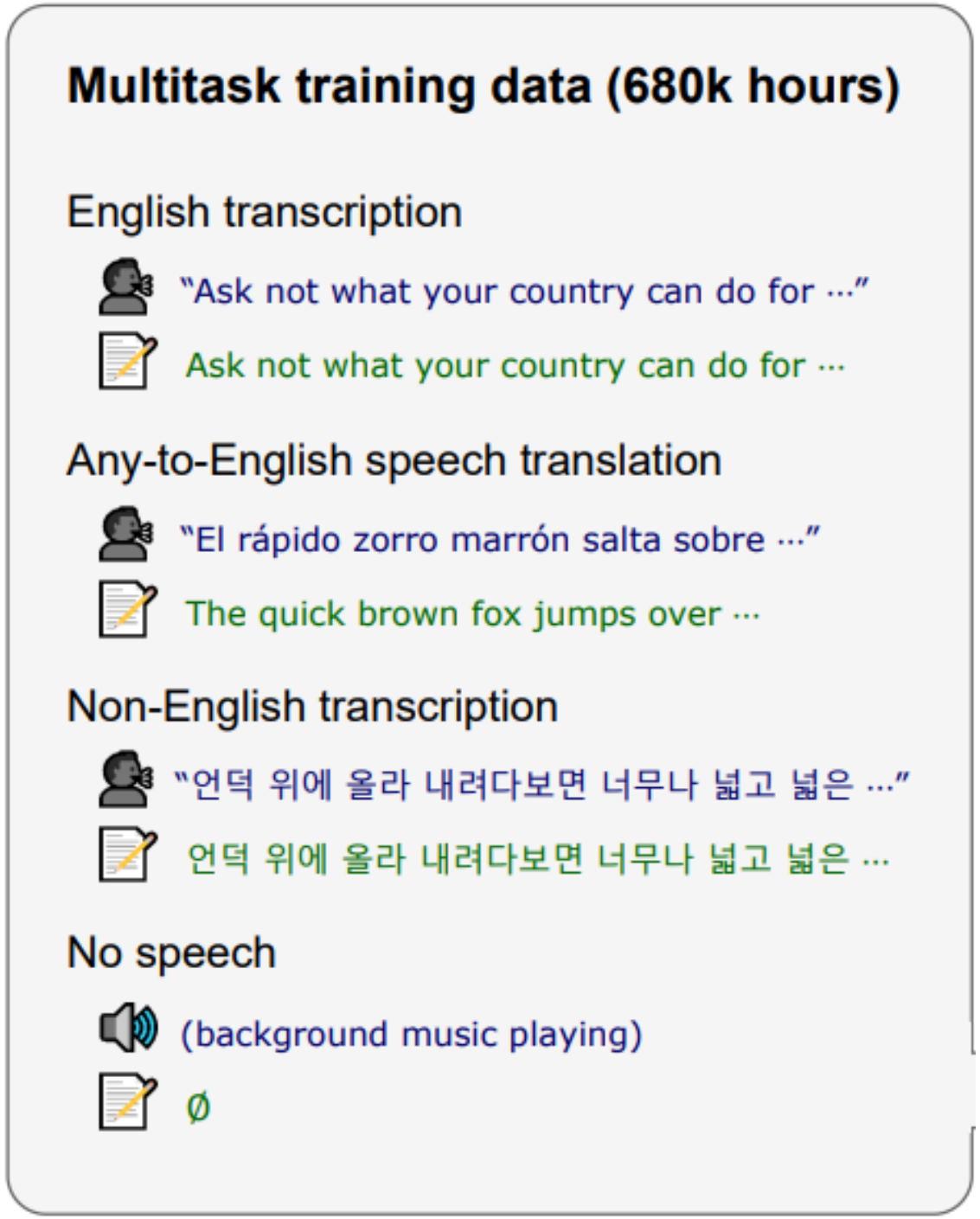
Access must be ***free of charge***.

Initial experiments with Liepa-3 dataset

What is Whisper?

An open-source neural model for Automatic Speech Recognition.

- ✓Developed by OpenAI.
- ✓Trained on ≈***680,000 hours*** of ***multilingual*** and ***multitask*** data.
- ✓Based on the Transformer architecture.
- ✓Freely available through Hugging Face and GitHub.



A model was trained on:

- ✓ multilingual speech recognition
- ✓ speech translation
- ✓ spoken language identification
- ✓ voice activity detection

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492-28518).

How the dataset used for training is structured?

Speech and text in English	438,218 h	65 %
Speech and text in 96 non-English languages	117,113 h	17 %
Non-English speech paired with English transcripts	125,739 h	18 %
Non-speech training segments used for VAD	*Not quantified in the dataset statistics*	

Language	Task	Hours
Lithuanian	Multilingual ASR	67 h
Lithuanian	Translation (LT→EN)	99 h
Polish	Multilingual ASR	4,278 h
Polish	Translation (PL→EN)	2,200 h

All data came from large, noisy web sources.

Whisper ASR performance (Common Voice 9 dataset)

Model	WER (Lithuanian)	WER (Polish)
Tiny	103.5 %	45.3 %
Base	91.3 %	32.8 %
Small	72.8 %	16.9 %
Medium	49.4 %	10.1 %
Large V2	43.9 %	9.0 %
Whisper large-v2	35.2 %	7.6 %

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023, July). Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning* (pp. 28492-28518).

Fine-Tuning Whisper for Lithuanian ASR

Corpus	Hours
Liepa3	2 098 h
Liepa2	948 h
Total training	3 046 h (692,000 segments, avg 16 s)
Test set	20.2 h (42,000 files, segments, avg 16 s)

Training Configuration

	Whisper Small	WhisperMedium
GPU	RTX 4060 M (8 GB)	RTX A5000 (24 GB)
per-device batch size	8	8
grad accumulation	2	2
Effective batch	16	16
Optimizer	AdamW	AdamW
Epochs	4	4

Results (WER %)

Model	WER (%)
Whisper Small FT	11.98 %
Whisper Medium FT	9.89 %

Thank you for your attention!

I'll be happy to answer any questions