Development of a Large Lithuanian Speech Corpus for Speech Recognition, Artificial Intelligence, and Other Innovative Language Technologies

Gediminas Navickas[1], Gailius Raškinis[2], Danguolė Mikulėnienė[3], Vytautas Kardelis[4], Indrė Makauskaitė[1], Pijus Kasparaitis[5], Margarita Beniušė[5], Laimonas Vėbra[1], Steponas Tolomanovas[1], Asta Kazlauskienė[2], Saulė Milčiuvienė[2], Gražina Korvel[1]

Background

The lack of robust and accessible speech resources for the Lithuanian language poses a challenge to its digitization, including efforts related to speech recognition, artificial intelligence (AI), and language technologies. This issue is acknowledged in the State Digitization Development Program 2021-2030 of the Ministry of the Economy and Innovation of the Republic of Lithuania. The program emphasizes the need to integrate advanced tools and technological solutions to improve the accessibility, security, and efficiency of e-services at both the national and international levels.

The Goal of Project

The project "**Development of the Large Lithuanian Speech Corpus (LIEPA-3)**" contributes to the digitization of the Lithuanian language. During the project, a large **corpus** of **10,000 hours of annotated speech** will be created.

Key features

- The corpus consists of Lithuanian speech recordings (non-Lithuanian language content limited to 0.1%).
- High quality, noise-varied recordings.
- Openly licensed and available on multiple platforms.

Corpus Diversity

- Balanced composition based on speaker gender, age, and regional background.
- Balanced phonetic coverage of Lithuanian sounds and thematic diversity.
- Stylistic diversity includes 5000 hours of read speech, 4900 hours of spontaneous speech.
- At least 100 hours of speech with jargon, and uncensored vocabulary.
- At least 100 hours dedicated to dialect-specific speech.

Annotation Requirements:

- All recordings will include synchronized utterance-level annotations
- At least 500 hours of recordings will be annotated at phoneme level.

Project implementation

Duration: **29 July 2024 – 30 April 2026**
Lead partner: **Vilnius University**
Partners: **Vytautas Magnus University**, **Institute of the Lithuanian Language**

Project webpage: http://LIEPA3.rastija.lt

Expected Results

The results of this project will facilitate innovation in AI and digital services, as well as improve public access to e-services in Lithuania. It will simplify interactions with digital platforms, promote digital inclusion and contribute to the broader adoption of AI technologies in various sectors. Ultimately, LIEPA-3 will support a more connected and digitally literate society by providing essential resources for developing user-friendly digital services. Also, the created speech corpus will be a good source for researchers in many fields, primarily contributing to linguistic research and informatics.

Authors Affiliations

- [1] Institute of Data Science and Digital Technologies, Vilnius University
- [2] Institute of Digital Resources and Interdisciplinary Research, Vytautas Magnus University
- [3] Institute of the Lithuanian Language
- [4] Institute of Applied Linguistics, Department of the Lithuanian Language, Vilnius University
- [5] Institute of Computer Science, Vilnius University

Funded by the European Union
NextGenerationEU
NAUJOS KARTOS LIETUVA

Contact: Gediminas Navickas, gediminas.navickas@mif.vu.lt