



# DAMSS

DATA ANALYSIS  
METHODS FOR SOFTWARE  
SYSTEMS



**15th Conference on**

# **DATA ANALYSIS METHODS for Software Systems**

**November 28–30, 2024**

Druskininkai, Lithuania, Hotel “Europa Royale”

<https://www.mii.lt/DAMSS>

**Co-Chairmen:**

Prof. **Gintautas Dzemyda** (Vilnius University, Lithuanian Academy of Sciences)

Dr. **Saulius Maskeliūnas** (Lithuanian Computer Society)

**Programme Committee:**

Dr. **Jolita Bernatavičienė** (Lithuania)

Prof. **Juris Borzovs** (Latvia)

Prof. **Janis Grundspenkis** (Latvia)

Prof. **Janusz Kacprzyk** (Poland)

Prof. **Ignacy Kaliszewski** (Poland)

Prof. **Bożena Kostek** (Poland)

Prof. **Tomas Krilavičius** (Lithuania)

Prof. **Olga Kurasova** (Lithuania)

Assoc. Prof. **Tatiana Tchemisova** (Portugal)

Assoc. Prof. **Gintautas Tamulevičius** (Lithuania)

Prof. **Julius Žilinskas** (Lithuania)

**Organizing Committee:**

Dr. **Jolita Bernatavičienė**

Prof. **Olga Kurasova**

Assoc. Prof. **Viktor Medvedev**

**Laima Paliulionienė**

Assoc. Prof. **Martynas Sabaliauskas**

Prof. **Povilas Treigys**

**Contacts:**

Dr. Jolita Bernatavičienė

*jolita.bernataviciene@mif.vu.lt*

Tel. (+370 5) 2109 315

Prof. Olga Kurasova

*olga.kurasova@mif.vu.lt*

Copyright © 2024 Authors. Published by Vilnius University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

<https://doi.org/10.15388/DAMSS.15.2024>

ISBN 978-609-07-1112-5 (digital PDF)

© Vilnius University, 2024

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

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

<sup>1</sup> Institute of Data Science and Digital Technologies, Vilnius University

<sup>2</sup> Institute of Digital Resources and Interdisciplinary Research, Vytautas Magnus University

<sup>3</sup> Institute of the Lithuanian Language

<sup>4</sup> Institute of Applied Linguistics, Department of the Lithuanian Language, Vilnius University

<sup>5</sup> Institute of Computer Science, Vilnius University

*gediminas.navickas@mif.vu.lt*

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 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. The sources will consist of 5000 hours of read speech, 4900 hours of spontaneous speech and 100 hours of four Lithuanian dialects. The corpus will be freely available in open formats on different platforms.

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.