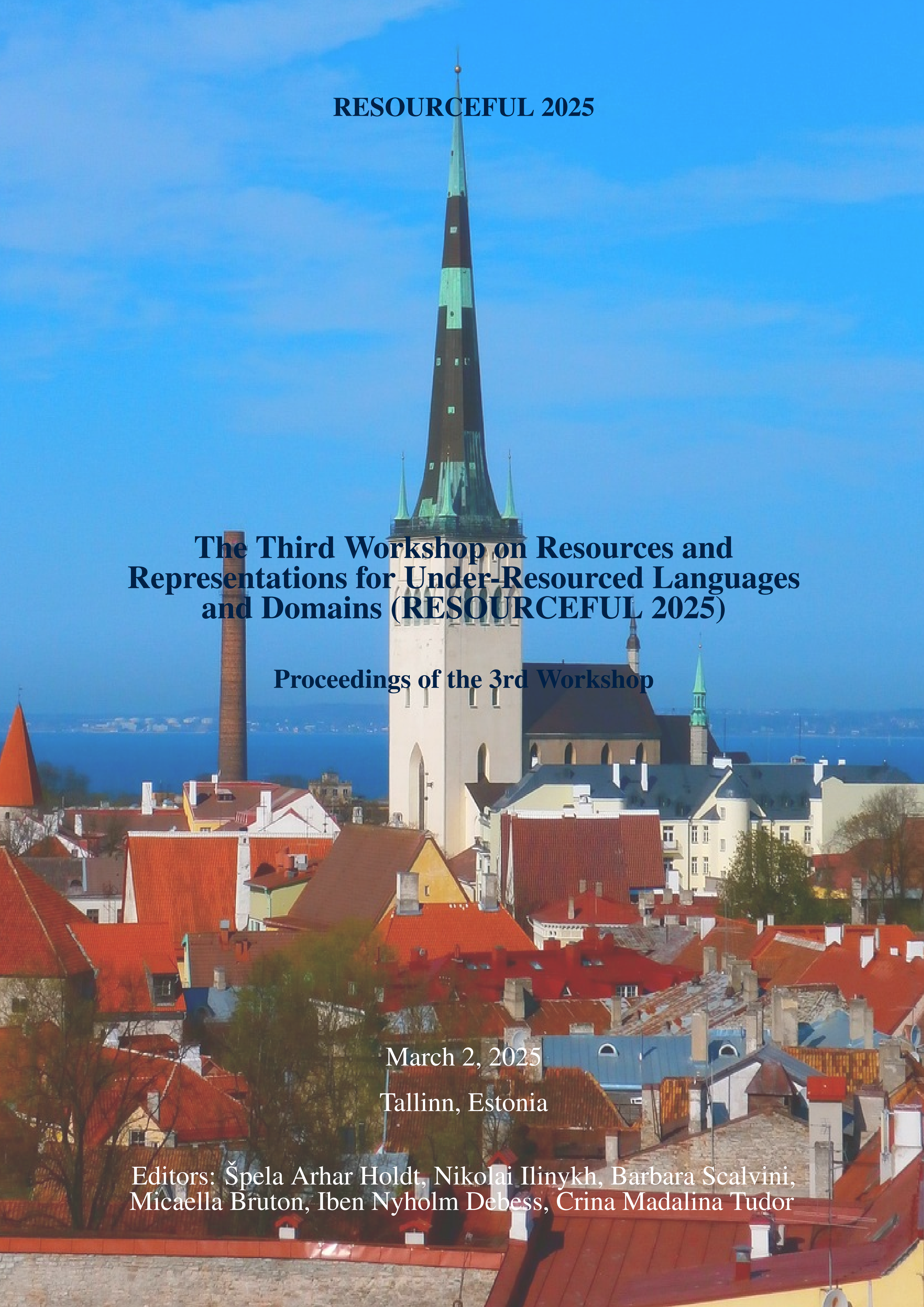

RESOURCEFUL 2025

The Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL 2025)

Proceedings of the 3rd Workshop

March 2, 2025

Tallinn, Estonia

Editors: Špela Arhar Holdt, Nikolai Ilinykh, Barbara Scalvini, Micaella Bruton, Iben Nyholm Debess, Crina Madalina Tudor

The RESOURCEFUL organisers gratefully acknowledge the support from the following organisations:

University of Ljubljana

UNIVERSITY OF
GOTHENBURG

UNIVERSITY OF
THE FAROE ISLANDS

Stockholm
University

UNIVERSITY OF HELSINKI

UNIVERSITY
OF OSLO

UPPSALA
UNIVERSITET

SPRÅKBANKENTEXT

Front-cover photo by avva from https://pixabay.com/photos/church-oleviste-churches-city-110716/

ISBN 978-9908-53-121-2

This work has been supported by Språkbanken – jointly funded by its 10 partner institutions and the Swedish Research Council (2025–2028; grant no. 2023-00161). It has also been supported by HUMINFRA with funding from the Swedish Research Council (grant no. 2023-00171).

Editors:
Špela Arhar Holdt, Nikolai Ilinykh, Barbara Scalvini, Micaella Bruton, Iben Nyholm Debess, Crina Madalina Tudor

Preface

The third workshop on resources and representations for under-resourced languages and domains was held in Tallinn, Estonia, on March 2nd, 2025. The workshop was conducted in person but also provided an option for online participation. In alignment with the goals of the previous two workshops in 2020 and 2023, RESOURCEFUL-2025 explored the role of resource type and quality available to computational linguists, as well as the challenges and directions for constructing new resources in light of the latest trends in natural language processing, computational linguistics, and artificial intelligence. The workshop provided a forum for discussions between the two communities involved in building data-driven and annotation-driven resources.

The call for papers for RESOURCEFUL-2025 requested work on the following topics:

- The types of linguistic knowledge that should be captured by models across different contexts and tasks
- Practical methods for sampling and extracting knowledge
- The relevance of traditional NLP resources for use in data-driven approaches
- The use of data-driven approaches to enhance expert-driven annotation processes
- Current challenges faced in expert-based annotation
- Crowdsourcing and citizen science initiatives to build and enrich linguistic resources
- Methods for evaluating and mitigating unwanted biases in linguistic models and data
- Creating anonymized and pseudonymized datasets and models
- Evaluating the role of modern LLMs in the creation of new linguistic resources

We invited both archival (long and short papers) and non-archival submissions. In total, 33 submissions were received, of which 23 were archival. The program committee (PC) consisted of 33 members (excluding 13 Program Chairs), who served as reviewers. Based on the PC assessments regarding the content and quality of the submissions, the program chairs decided to accept 26 submissions for presentation and publication. Together with the 4 non-archival submissions, we devised a program consisting of 7 talks and 17 posters. The accepted submissions covered topics related to working with specific linguistic characteristics, investigating and analyzing specific aspects of languages or contexts, and exploiting methods for analyzing, exploring, and improving the quality and quantity of low-resourced and medium-resourced languages, domains, and applications. The topics presented in the accepted submissions led to the emergence of the following themes and questions for the panel discussion:

1. **Analysis and Exploration of Linguistic Characteristics and Features in Specific Languages**. This line of presented work focused on the linguistic characteristics and features of various languages, including Uzbek, Korean, Haitian Creole, Central Australian languages, Faroese, Spanish, Icelandic, Ottoman Turkish, Brazilian Indigenous languages, Latvian, Niger-Congo languages, Swedish, Finnish, Armenian, German, Slovene, Luxembourgish dialects, Kirundi, Komi-Permyak, Komi-Zyrian, Polish, and English.

   - What are the current challenges in expert-based annotation, and how can data-driven approaches facilitate this process?
   - Which resources and corpora, and in which modalities (text, image, video, audio), are missing for the computational modeling of the aforementioned languages?

- What are the real-world problem domains where these corpora and models can be applied, such as healthcare or cultural preservation?

2. **Development and Evaluation of Datasets and Models for Linguistic Analysis and NLP Tasks**. Questions regarding datasets include, but are not limited to, the creation of UD treebanks, phonotactic corpora, and various annotation tools for speakers of Indigenous languages. Tasks encompass part-of-speech tagging, linguistic variation, code-switching, OCR error correction, question-answering, noun classification, annotation of political attitudes, text generation, personal information detection, and benchmarking with large language models.

   - What strategies can be implemented to improve specific tasks with no available training data?
   - How can we ensure fairness and inclusivity in NLP models and datasets?
   - How can we assess biases in created datasets and inform users about them?

3. **Challenges in Expert-Based Annotation vs. Data-Driven Approaches**.

   - What are the current bottlenecks in expert-based annotation, and where do data-driven, semi-supervised, or active learning methods offer improvements?
   - Can hybrid approaches be developed to leverage the strengths of both human expertise and automated techniques?
   - How can we standardize annotation practices to improve cross-dataset compatibility and overall quality?

Completing the program were three invited keynote speakers: Beáta Megyesi from Stockholm University, Jussi Karlgren from Silo AI and Joshua Wilbur from University of Tartu.

Words of appreciation and acknowledgment are due to the program committee, the local NoDaLiDa/Baltic-HLT 2025 organisers, and OpenReview.

The RESOURCEFUL 2025 Program Chairs

Organizing Committee

Organizing Committee

Špela Arhar Holdt, University of Ljubljana, Slovenia
Nikolai Ilinykh, CLASP, University of Gothenburg, Sweden
Barbara Scalvini, University of the Faroe Islands, Faroe Islands
Mattias Appelgren, University of Gothenburg, Sweden
Micaella Bruton, Stockholm University, Sweden
Dana Dannélls, Språkbanken Text, University of Gothenburg, Sweden
Simon Dobnik, CLASP, University of Gothenburg, Sweden
Crina Tudor, Stockholm University, Sweden
Joakim Nivre, RISE and Uppsala University, Sweden
Iben Nyholm Debess, University of the Faroe Islands, Faroe Islands
Sara Stymne, Uppsala University, Sweden
Jörg Tiedemann, University of Helsinki, Finland
Lilja Øvrelid, University of Oslo, Norway

Program Committee

Program Chairs

Mattias Appelgren, Göteborg University
Micaella Bruton, Stockholm University
Dana Dannélls, Göteborg University
Iben Nyholm Debess, University of the Faroe Islands, Faroe Islands
Simon Dobnik, University of Gothenburg
Špela Arhar Holdt, University of Ljubljana
Nikolai Ilinykh, Göteborg University
Joakim Nivre, Uppsala University
Barbara Scalvini, University of the Faroe Islands
Sara Stymne, Uppsala University
Jörg Tiedemann, University of Helsinki
Crina Tudor, Stockholm University
Lilja Øvrelid, Dept. of Informatics, University of Oslo

Reviewers

David Alfter

Micaella Bruton

Peter Ebert Christensen

Dana Dannélls, Simon Dobnik, Luise Dürlich

Emilie Marie Carreau Francis

Evangelia Gogoulou

Carlos Daniel Hernández Mena, Špela Arhar Holdt

Nikolai Ilinykh

Herbert Lange, Staffan Larsson, Ying Li, Ellinor Lindqvist

Arianna Masciolini, John Philip McCrae, Felix Morger, Amanda Muscat, Ricardo Muñoz Sánchez, Petter Mæhlum

Joakim Nivre, Bill Noble

Robert Östling, Lilja Øvrelid

Danila Petrelli

Michael Rießler

Annika Simonsen, Maria Irena Szawerna

Jörg Tiedemann, Tiago Timponi Torrent, Crina Tudor

Thomas Vakili

Table of Contents