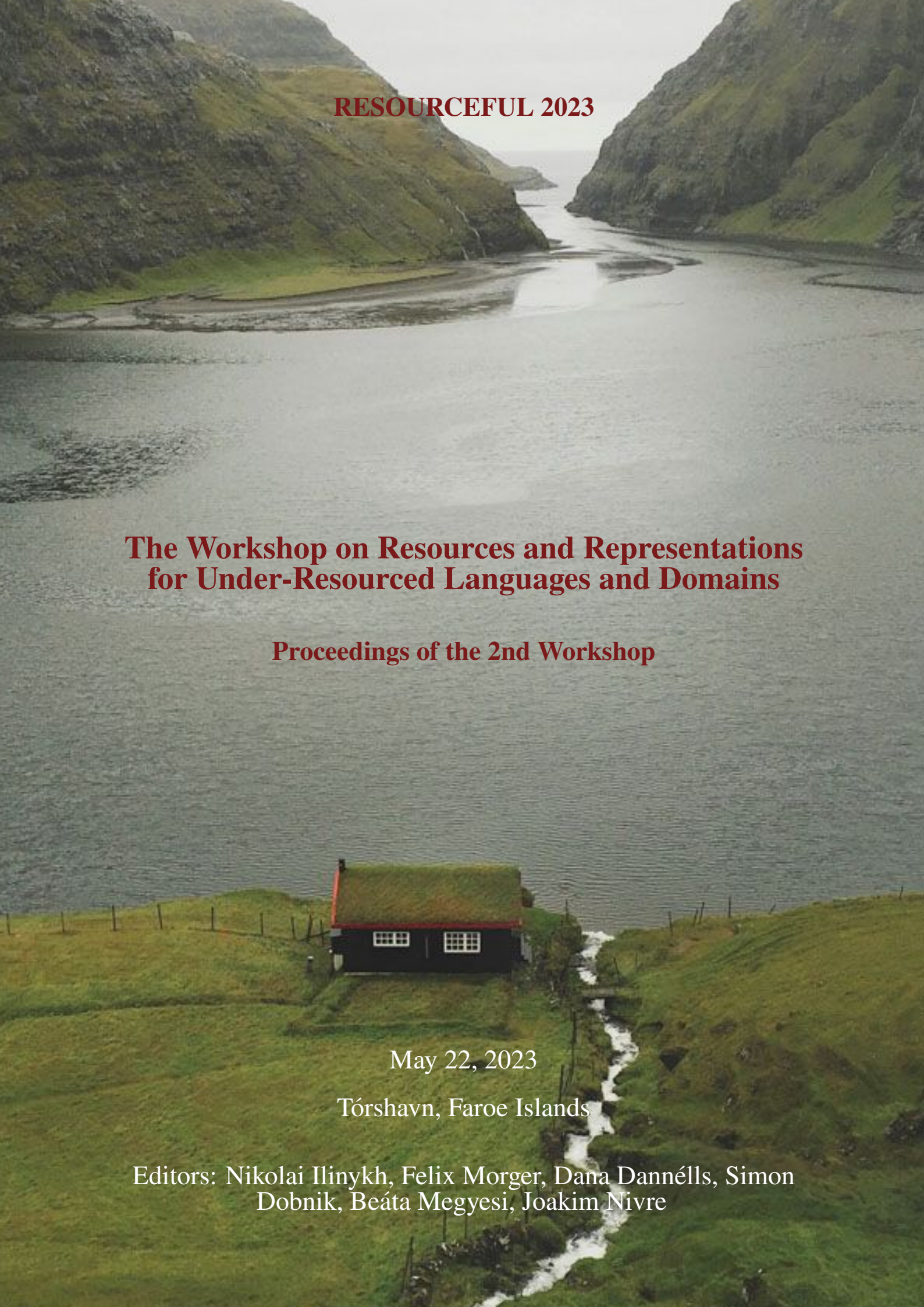

RESOURCEFUL 2023

The Workshop on Resources and Representations for Under-Resourced Languages and Domains

Proceedings of the 2nd Workshop

May 22, 2023

Tórshavn, Faroe Islands

Editors: Nikolai Ilinykh, Felix Morger, Dana Dannélls, Simon Dobnik, Beáta Megyesi, Joakim Nivre

Partner organisations

UNIVERSITY OF GOTHENBURG

CLASP centre for linguistic theory and studies in probability

UPPSALA UNIVERSITET

UNIVERSITY OF OSLO

AI SWEDEN

UNIVERSITY OF HELSINKI

SPRÅKBANKENTEXT

Front-cover photo: Sage Stephens(@storyofsage, https://www.instagram.com/p/BZOuVK8HkUO/)

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 978-1-959429-73-9

Editors:
Nikolai Ilinykh, Felix Morger, Dana Dannélls, Simon Dobnik, Beáta Megyesi, Joakim Nivre

Preface

The second workshop on resources and representations for under-resourced language and domains was held in Tórshavn, Faroe Islands on May 22nd, 2023. The workshop was conducted in a physical setting, allowing for potential hybrid participation.

Continuing with the aim of the first edition in 2020, RESOURCEFUL explored the role of the kind and the quality of resources that are available to us, as well as the challenges and directions for constructing new resources in light of the latest trends in natural language processing. The workshop has provided a forum for discussions between the two communities involved in building data-driven and annotation-driven resources.

Data-driven machine-learning techniques in natural language processing have achieved remarkable performance (e.g., BERT, GPT, ChatGPT) but the secret of their success is large quantities of quality data (which is mostly text). Interpretability studies of large language models in both text-only and multi-modal setups have revealed that even in cases where large text datasets are available, the models still do not cover all the contexts of human social activity and are prone to capturing unwanted bias where data is focused towards only some contexts. A question has also been raised whether textual data is enough to capture semantics of natural language processing, or whether other modalities such as visual representations or a situated context of a robot might be required. Annotation-driven resources have been constructed over years based on theoretical work in linguistics, psychology and related fields and a large amount of work has been done both theoretically and practically. They are valuable for evaluating data-driven resources and for improving their performances.

The call for papers for RESOURCEFUL 2023 requested work on resource creation, representation learning and interpretability in data-driven and expert-driven machine learning setups and both uni-modal and multi-modal scenarios. We invited both archival (long and short papers) and non-archival submissions.

In the call for papers we invited students, researchers, and experts to address and discuss the following questions:

- What is relevant linguistic knowledge the models should capture and how can this knowledge be sampled and extracted in practice?
- What kind of linguistic knowledge do we want and can capture in different contexts and tasks?
- To what degree are resources that have been traditionally aimed at rule-based natural language processing approaches relevant today both for machine learning techniques and hybrid approaches?
- How can they be adapted for data-driven approaches?
- To what degree data-driven approaches can be used to facilitate expert-driven annotation?
- What are current challenges for expert-based annotation?
- How can crowd-sourcing and citizen science be used in building resources?
- How can we evaluate and reduce unwanted biases?

In total **21** submissions were received of which **19** were archival submissions. The programme committee (PC) consisted of 21 members (excluding the 12 program chairs), who worked as reviewers. Based on the PC assessments regarding the content, and quality of the submissions, the program chairs decided to accept **16** submissions for presentation and publication. Together with the 2 non-archival submissions we

devised a programme consisting of *8* talks and *10* posters. The accepted submissions covered topics about working with specific linguistic characteristics, investigating and analysing specific aspects of languages or contexts, exploiting methods for analysing, and exploring and improving the quality and quantity of low-resourced and medium-resourced languages, domains and applications.

The topics presented in the accepted submissions resulted in the emergence of the following themes and questions for the panel discussion:

1. Exploration of linguistic characteristics and features of language-related topics in specific languages (Danish, Norwegian, Okinawan, Sakha, Sign Language, Spanish, Swedish, and Uralic)
    - What are the current challenges for expert-based annotation, and how can data-driven approaches facilitate this process?
2. Development of datasets or models for linguistic analysis and NLP tasks (automatic speech recognition, corpus and lexicon construction, language transfer, parallel annotations, part-of-speech tagging, prediction/generation of gaze in conversation, sentiment and negation modelling, and word substitution)
    - What strategies can be implemented to improve specific tasks with no training data available? How can we ensure fairness and inclusivity?
    - How can we assess the biases present in created datasets, and inform users about these biases?
3. Understanding the impact of technologies and resources within specific contexts (English and multilingual models for Swedish, parallel annotations for European languages, and universal dependencies treebanking)
    - How can multilingual approaches be effectively utilised in building linguistic resources, and what are their contributions to resource development?

Completing the programme are two invited keynote speakers with a strong connection to, on the one hand, data-driven methods by Jörg Tiedemann (University of Helsinki, Finland) and, on the other hand, expert-driven annotations by Darja Fišer (Institute of Contemporary History, Ljubljana, Slovenia).

Words of appreciation and acknowledgment are due to the program committee, the local NoDaLiDa organisers, and the OpenReview.

The RESOURCEFUL program chairs

Organizing Committee

Program Chairs

Dana Dannélls, Språkbanken Text, University of Gothenburg, Sweden
Simon Dobnik, CLASP, University of Gothenburg, Sweden
Adam Ek, CLASP, University of Gothenburg, Sweden
Stella Frank, University of Copenhagen, Denmark
Nikolai Ilinykh, CLASP, University of Gothenburg, Sweden
Beáta Megyesi, Uppsala University, Sweden
Felix Morger, Språkbanken Text, University of Gothenburg, Sweden
Joakim Nivre, RISE and Uppsala University, Sweden
Magnus Sahlgren, AI Sweden, Sweden
Sara Stymne, Uppsala University, Sweden
Jörg Tiedemann, University of Helsinki, Finland
Lilja Øvrelid, University of Oslo, Norway

Program Committee

Reviewers

Manex Agirrezabal

Meriem Beloucif

Robin Cooper, Amaru Cuba Gyllensten

Luise Dürlich

Desmond Elliott

Markus Forsberg

Evangelia Gogoulou

Sardana Ivanova

Richard Johansson

Ellinor Lindqvist

Petter Mæhlum

Bill Noble

Robert Östling

Bolette S. Pedersen, Danila Petrelli, Miriam R. L. Petruck, Eva Pettersson

Simeon Schüz, Per Erik Solberg

Thomas Vakili

Keynote Talk: Democratizing Machine Translation with OPUS and OPUS-MT

Jörg Tiedemann
University of Helsinki, Helsinki, Finland

Abstract: The demand for translation is ever growing and this trend will not stop. Being able to access the same kind of information is a fundamental prerequisite for equality in society and translation plays a crucial role when fighting discrimination based on language barriers. Efficient tools and a better coverage of the linguistic diversity in the World are necessary to cope with the amount of material that needs to be handled. Our mission is to support the development of high quality tools for automatic and computer-assisted translation by providing open services and resources that are independent of commercial interests and profit-driven companies. Equal information access is a human right and not only a privilege for people who can pay for it. In this talk I will discuss the current state of OPUS-MT, our project on open neural machine translation and the challenges that we try to tackle with multilingual NLP, transfer learning and data augmentation. I will report about on-going work on knowledge distillation, the creation of compact models for real-time translation and our work on modularization of neural MT.

Bio: Jörg Tiedemann is professor of language technology at the Department of Digital Humanities at the University of Helsinki. He received his PhD in computational linguistics for work on bitext alignment and machine translation from Uppsala University before moving to the University of Groningen for 5 years of post-doctoral research on question answering and information extraction. His main research interests are connected with massively multilingual data sets and data-driven natural language processing and he currently runs an ERC-funded project on representation learning and natural language understanding.

Keynote Talk: The Role of the CLARIN Research Infrastructure in the Era of Data-Intensive Language Studies

Darja Fišer
Institute of Contemporary History, Ljubljana, Slovenia

Abstract: Advances in digitization and datafication have been transformative for linguistics and other disciplines that work with language materials. This has increased the need for research infrastructures that supports the development, documentation, archiving, dissemination, reuse and citation of language resources and tools which is prerequisite for verifiable and reproducible research. In this talk I will present the recent achievements and ongoing work of the CLARIN research infrastructure which is based on the Open Science paradigm and FAIR data principles. It provides easy and sustainable access to digital language data and offers advanced tools to discover, explore, annotate, analyse, and combine such datasets, wherever they are located. This is enabled through a networked federation of centres: language data repositories, service centres, and knowledge centres with single sign-on access for all members of the academic community in all participating countries. Tools, data and metadata from different centres are interoperable so that data collections can be combined and tools from different sources can be chained to perform operations at different levels of complexity.

Bio: Darja Fišer is Executive Director of CLARIN. She has a background in corpus linguistics and language resource creation. She has been Associate Professor at the Faculty of Arts, University of Ljubljana, since 2019, Senior Research Fellow at the Institute of Contemporary History since 2021, and is leading the new national research programme for Digital Humanities in Slovenia. She is also serving as a member of the Scientific Advisory Board of the Austrian Centre for Digital Humanities at the Austrian Academy of Sciences, the National Interdisciplinary Research E-Infrastructure for Bulgarian Language and Cultural Heritage Resources and Technologies, and the Czech National Corpus research infrastructure of the Institute of the Czech National Corpus at Charles University.

Table of Contents