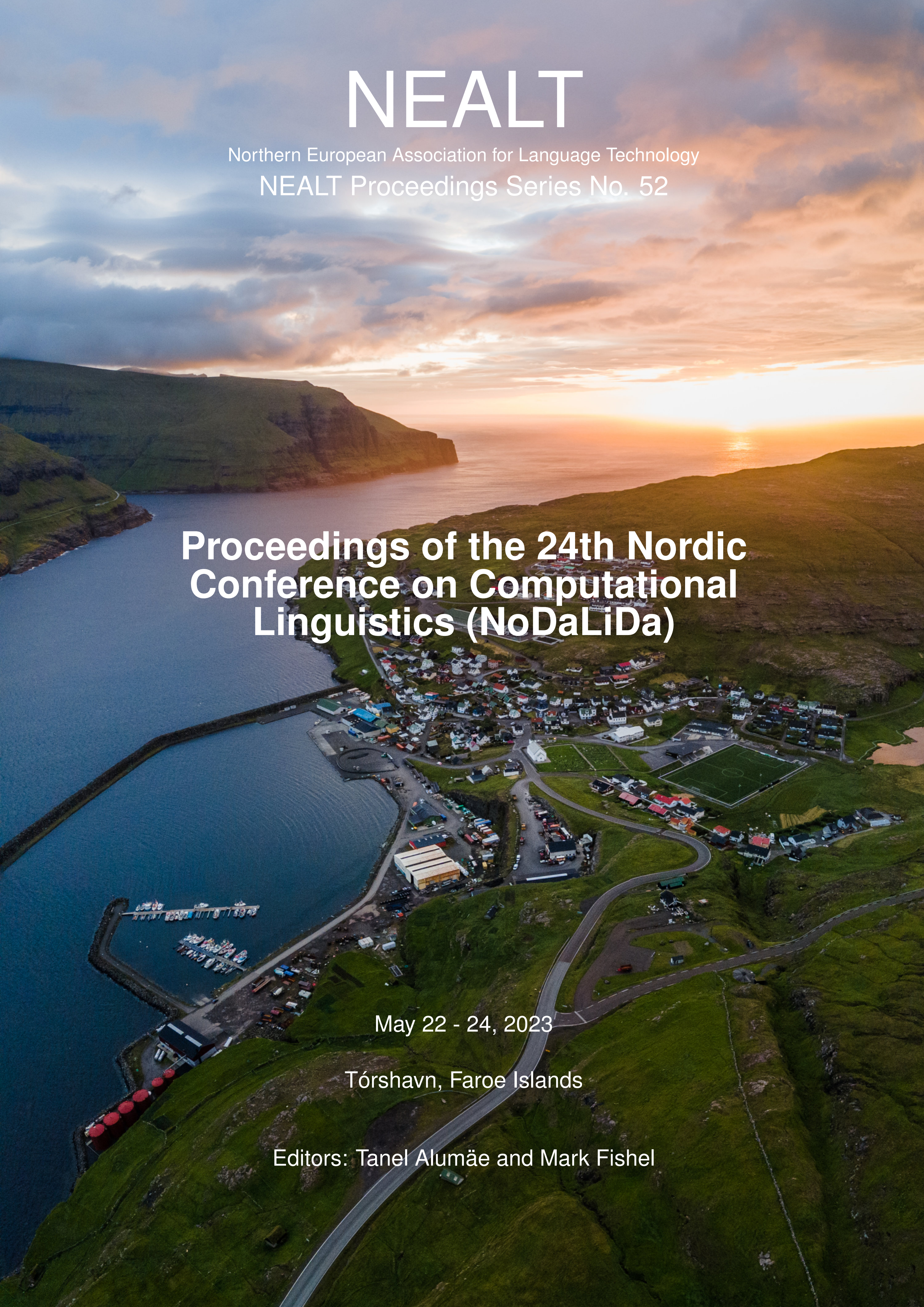

NEALT

Northern European Association for Language Technology

NEALT Proceedings Series No. 52

Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)

May 22 - 24, 2023

Tórshavn, Faroe Islands

Editors: Tanel Alumäe and Mark Fishel

NoDaLiDa 2023

24th Nordic Conference on Computational Linguistics (NoDaLiDa)

Proceedings of the Conference

May 22-24, 2023

The NoDaLiDa organizers gratefully acknowledge the support from the following sponsors.

Silver

DICTUS

elektron

NATIONELLA
SPRÅKBANKEN

Bronze

formula | BANKNORDIK

Other

MÁLRÁÐIÐ

DIGITALISERINGSSTYRELSEN

NEALT
NORTHERN EUROPEAN ASSOCIATION FOR LANGUAGE TECHNOLOGY

TÓRSHAVNAR KOMMUNA

UNIVERSITY OF
THE FAROE ISLANDS

©2023 University of Tartu Library

Front-cover photo: Luca Renner (@lucarennerphotography, http://visitfaroeislands.com)

Published by:

University of Tartu Library, Estonia
NEALT Proceedings Series, No. 52
Indexed in the ACL Anthology

ISBN: 978-99-1621-999-7
ISSN: 1736-8197 (Print)
ISSN: 1736-6305 (Online)

Volume Editors:
Tanel Alumäe and Mark Fishel

Message from the General Chair

It is my great pleasure and honor to welcome you to the 24th Nordic Conference on Computational Linguistics (NoDaLiDa 2023)!

After a couple of years' worth of conferences cancelled or held online (including the previous NoDaLiDa) we are extremely happy that NoDaLiDa 2023 is an onsite event. This is especially exciting given that for the first time in the history of NoDaLiDa conferences it takes place in Tórshavn, Faroe Islands.

The conference features three types of papers: long, short and demo papers. We are truly grateful to all the authors of papers submitted to this year's conference, with 130 papers submitted, a more than 40% increase over last year's yield! In total, we accepted 79 papers: 49 long papers, 26 short papers and 4 demo papers. More than half of the accepted papers are student papers, in which the first author is a student (29 long, 17 short and 2 demo papers). We would like to thank the 113 members of the program committee who reviewed the papers for their contributions!

The 79 accepted papers are grouped into 12 oral and 2 poster sessions. In addition to these regular sessions the conference program also includes three keynote talks. We would like to extend our gratitude to the keynote speakers for agreeing to present their work at NoDaLiDa. Georg Rehm from DFKI will talk on the topic of "Towards Digital Language Equality in Europe: An Overview of Recent Developments". Hjalmar P. Petersen will talk about "Aspects of the structure of Faroese". Marta R. Costa-Jussà from Meta will talk about "No-language-left-behind: Scaling Human-Centered Machine Translation and Toxicity at Scale".

The main conference program is preceded by three workshops: NLP for Computer-Assisted Language Learning (NLP4CALL), the Constraint Grammar Workshop and Resources and representations for under-resourced languages and domains (RESOURCEFUL'2023). We thank the workshop organizers for their efforts and for expanding the main conference program with a focus on more specific research topics.

I would like to express sincere gratitude to the entire team behind organizing NoDaLiDa 2023. I was honored to receive the invitation to serve as the general chair from the NEALT board; thank you for trusting me with this role. My deepest gratitude goes to Tanel Alumäe for serving as the publications chair and his active participation, Inguna Skadiņa for serving as the workshop chair as well as Iben Nyholm Debess for serving as the main local chair and smoothly handling all associated aspects of conference organization. I also want to thank the rest of the program chairs, Lilja Øvrelid and Christian Hardmeier and the local co-chairs Bergur Djurhuus Hansen, Peter Juel Henrichsen and Sandra Saxov Lamhauge. Thank you everyone for your contributions, you are awesome!

NoDaLiDa 2023 received financial support from several institutions and we would like to thank them here: NEALT, Dictus, Málráðið, Tórshavnar kommuna, BankNordik, Digitaliseringsstyrelsen, University of the Faroe Islands, Nationella språkbanken, Elektron and Formula.

Welcome and enjoy the 24th Nordic Conference on Computational Linguistics!

Mark Fishel, General Chair

Tartu

May 2023

Organizing Committee

General Chair

Mark Fishel, University of Tartu

Program Chairs

Tanel Alumäe, Tallinn University of Technology (Publication Chair)
Inguna Skadina, University of Latvia (Workshop Chair)
Christian Hardmeier, IT University of Copenhagen
Lilja Øvrelid, University of Oslo

Local Chair

Iben Nyholm Debess, University of the Faroe Islands

Local Co-Chairs

Bergur Djurhuus Hansen, University of the Faroe Islands
Peter Juel Henrichsen, Danish Language Council
Sandra Saxov Lamhauge, University of the Faroe Islands

Reviewers

Yvonne Adesam, Lars Ahrenberg, David Alfter, Krasimir Angelov, Ilze Auzina

Eduard Barbu, Jeremy Barnes, Valerio Basile, Ali Basirat, Timo Baumann, Jean-Philippe Bernardy, Arianna Bisazza, Jari Björne, Sidsel Boldsen, Gerlof Bouma, Gosse Bouma, Johan Boye, Chloé Braud, Maja Buljan

Marie Candito, Lin Chen, Mathias Creutz

Hercules Dalianis, Dana Dannélls, Leon Derczynski, Stefanie Dipper, Simon Dobnik

Adam Ek

Antske Fokkens

Filip Ginter, Rob Van Der Goot, Normunds Gruzitis, Tamás Grósz, Jon Gudnason

Mareike Hartmann, Daniel Hershcovich

Tommi Jauhiainen, Richard Johansson

Heiki-Jaan Kaalep, Kaarel Kaljurand, Katharina Kann, Jurgita Kapočiūtė-Dzikienė, Jussi Karlgren, Andre Kasen, Andreas Kirkedal, Roman Klinger, Mare Koit, Jiaming Kong, Marco Kuhlmann, Mikko Kurimo, Robin Kurtz, Andrey Kutuzov

Ekaterina Lapshinova-Koltunski, Krister Lindén, Pierre Lison, Hrafn Loftsson, Jan Tore Lønning

Bruno Martins, Farrokh Mehryary, Einar Meister, Hans Moen, Kadri Muischnek, Kaili Müürisep

Costanza Navarretta, Anna Björk Nikulasdottir, Joakim Nivre, Pierre Nugues, Arild Brandrud Næss

Heili Orav, Robert Östling

Patrizia Paggio, Anthi Papadopoulou, Eva Pettersson, Mārcis Pinnis, Barbara Plank, Sampo Pyysalo

Alessandro Raganato, Liisa Rätsep

Magnus Sahlgren, Askars Salimbajevs, Baiba Valkovska Saulīte, Yves Scherrer, Miikka Silfverberg, Kairit Sirts, Raivis Skadiņš, Maria Skeppstedt, Steinþór Steingrímsson, Sara Stymne, Torbjørn Svendsen, Rune Sætre

Jörg Tiedemann, Samia Touileb, Trond Trosterud, Andre Tättar

Martti Vainio, Daniel Varab, Martin Volk, Elena Volodina

Fredrik Wahlberg

Roman Yangarber

Niklas Zechner, Heike Zinsmeister

Invited Talk: Towards Digital Language Equality in Europe: An Overview of Recent Developments

Georg Rehm
German Research Center for Artifical Intelligence

Digital Language Equality (DLE) “is the state of affairs in which all languages have the technological support and situational context necessary for them to continue to exist and to prosper as living languages in the digital age”, as we specified in one of our key reports of the EU project European Language Equality (ELE). Our empirical findings suggest that Europe is currently very far from having a situation in which all our languages are supported equally well through technologies. In this presentation, I’ll give an overview of the two ELE projects and their main results and findings with a special focus on the Nordic languages (including insights from the FSTP projects supported through ELE2). This will also include a brief look back into the past, especially discussing the question if and where we have seen progress in the last, say, 15 years. Furthermore, I’ll present an overview of our main strategic recommendations towards the European Union in terms of bringing about DLE in Europe by 2030. The presentation will conclude with a look at other relevant activities in Europe, including, critically the Common European Language Data Space project, which started in early 2023.

Invited Talk: No-language-left-behind: Scaling Human-Centered Machine Translation and Toxicity at Scale

Marta R. Costa-jussà
Meta AI

Machine Translation systems can produce different types of errors, some of which are characterized as critical or catastrophic due to the specific negative impact that they can have on users. In this talk, we focus on one type of critical error: added toxicity. We evaluate and analyze added toxicity in the context of NLLB-200 that open-sources models capable of delivering evaluated, high-quality translations directly between 200 languages. An automatic toxicity evaluation shows that added toxicity across languages varies from 0% to 5%. The output languages with the most added toxicity tend to be low-resource ones, and the demographic axes with the most added toxicity include sexual orientation, gender and sex, and ability. Making use of the input attributions allows us to further explain toxicity and our recommendations to reduce added toxicity are to curate training data to avoid mistranslations, mitigate hallucination and check unstable translations.

Invited Talk: Aspects of the Structure of Faroese

Hjalmar P. Petersen
University of Faroe Islands

Phonological changes and later morphologization have led to different complex alternations in Faroese. These are argued to emerge especially in small languages, with little contact and tight networks. The alternations will be exemplified with 'skerping', palatalization, glide insertion and the quantity-shift. There will be a discussion of the morphology-phonology interface, where the suggestion is that Faroese has 3 strata, stem1, stem2 and a word- strata. Syntactic variation and different construction will be addressed and illustrated; in this context reflexives are included and the present reorganization of the case system of complements of prepositions, where speakers use semantic and structural case in a certain way.

Table of Contents

Automated Claim Detection for Fact-checking: A Case Study using Norwegian Pre-trained Language Models
Ghazaal Sheikhi, Samia Touileb and Sohail Ahmed Khan 1

Evaluating the Impact of Text De-Identification on Downstream NLP Tasks
Cedric Lothritz, Bertrand Lebichot, Kevin Allix, Saad Ezzini, Tegawendé F. Bissyandé, Jacques Klein, Andrey Boytsov, Clément Lefebvre and Anne Goujon 10

Abstractive Text Summarization for Icelandic
Þór Sverrisson and Hafsteinn Einarsson 17

ASR Language Resources for Faroese
Carlos Daniel Hernández Mena, Annika Simonsen and Jon Gudnason 32

Good Reads and Easy Novels: Readability and Literary Quality in a Corpus of US-published Fiction
Yuri Bizzoni, Pascale Feldkamp Moreira, Nicole Dwenger, Ida Marie S. Lassen, Mads Rosendahl Thomsen and Kristoffer L. Nielbo 42

Detection and attribution of quotes in Finnish news media: BERT vs. rule-based approach
Maciej Janicki, Antti Kanner and Eetu Mäkelä 52

Dyslexia Prediction from Natural Reading of Danish Texts
Marina Björnsdóttir, Nora Hollenstein and Maria Barrett 60

Is Part-of-Speech Tagging a Solved Problem for Icelandic?
Örvar Kárason and Hrafn Loftsson 71

Multi-CrossRE A Multi-Lingual Multi-Domain Dataset for Relation Extraction
Elisa Bassignana, Filip Ginter, Sampo Pyysalo, Rob Van Der Goot and Barbara Plank 80

Microservices at Your Service: Bridging the Gap between NLP Research and Industry
Tiina Lindh-Knuutila, Hrafn Loftsson, Pedro Alonso Doval, Sebastian Andersson, Bjarni Barkarson, Héctor Cerezo-Costas, Jon Gudnason, Jökull Snær Gylfason, Jarmo Hemminki and Heiki-Jaan Kaalep 86

Slaapte or Sliep? Extending Neural-Network Simulations of English Past Tense Learning to Dutch and German
Xiulin Yang, Jingyan Chen, Arjan van Eerden, Ahnaf Mozib Samin and Arianna Bisazza 92

Class Explanations: the Role of Domain-Specific Content and Stop Words
Denitsa Saynova, Bastiaan Bruinsma, Moa Johansson and Richard Johansson 103

Constructing Pseudo-parallel Swedish Sentence Corpora for Automatic Text Simplification
Daniel Holmer and Evelina Rennes 113

Who said what? Speaker Identification from Anonymous Minutes of Meetings
Daniel Holmer, Lars Ahrenberg, Julius Monsen, Arne Jönsson, Mikael Apel and Marianna Blix Grimaldi 124

On the Concept of Resource-Efficiency in NLP
Luise Dürlich, Evangelia Gogoulou and Joakim Nivre 135

Identifying Token-Level Dialectal Features in Social Media
Jeremy Barnes, Samia Touileb, Petter Mæhlum and Pierre Lison . . . 146

NorQuAD: Norwegian Question Answering Dataset
Sardana Ivanova, Fredrik Aas Andreassen, Matias Jentoft, Sondre Wold and Lilja Øvrelid . . . 159

Extracting Sign Language Articulation from Videos with MediaPipe
Carl Börstell . . . 169

Named Entity layer in Estonian UD treebanks
Kadri Muischnek and Kaili Müürisep . . . 179

ScandEval: A Benchmark for Scandinavian Natural Language Processing
Dan Saattrup Nielsen . . . 185

BRENT: Bidirectional Retrieval Enhanced Norwegian Transformer
Lucas Georges Gabriel Charpentier, Sondre Wold, David Samuel and Egil Rønningstad . . . 202

Machine vs. Human: Exploring Syntax and Lexicon in German Translations, with a Spotlight on Anglicisms
Anastassia Shaitarova, Anne Göhring and Martin Volk . . . 215

Training and Evaluating Norwegian Sentence Embedding Models
Bernt Ivar Utstøl Nødland . . . 228

Dozens of Translation Directions or Millions of Shared Parameters? Comparing Two Types of Multilinguality in Modular Machine Translation
Michele Boggia, Stig-Arne Grönroos, Niki Andreas Loppi, Timothee Mickus, Alessandro Raganato, Jörg Tiedemann and Raúl Vázquez . . . 238

DanSumT5: Automatic Abstractive Summarization for Danish
Sara Kolding, Katrine Nymann, Ida Bang Hansen, Kenneth C. Enevoldsen and Ross Deans Kristensen-McLachlan . . . 248

CaptainA - A mobile app for practising Finnish pronunciation
Nhan Phan, Tamás Grósz and Mikko Kurimo . . . 265

DanTok: Domain Beats Language for Danish Social Media POS Tagging
Kia Kirstein Hansen, Maria Barrett, Max Müller-Eberstein, Cathrine Damgaard, Trine Naja Eriksen and Rob Van Der Goot . . . 271

Comparison of Current Approaches to Lemmatization: A Case Study in Estonian
Aleksei Dorkin and Kairit Sirts . . . 280

Generating Errors: OCR Post-Processing for Icelandic
Atli Jasonarson, Steinþór Steingrímsson, Einar Freyr Sigurðsson, Árni Davíð Magnússon and Finnur Ágúst Ingimundarson . . . 286

Generation of Replacement Options in Text Sanitization
Annika Willoch Olstad, Anthi Papadopoulou and Pierre Lison . . . 292

MeDa-BERT: A medical Danish pretrained transformer model
Jannik Skyttegaard Pedersen, Martin Sundahl Laursen, Pernille Just Vinholt and Thiusius Rajeeth Savarimuthu . . . 301

Standardising Pronunciation for a Grapheme-to-Phoneme Converter for Faroese
Sandra Saxov Lamhauge, Iben Nyholm Debess, Carlos Daniel Hernández Mena, Annika Simonsen and Jon Gudnason . . . 308

Using Membership Inference Attacks to Evaluate Privacy-Preserving Language Modeling Fails for Pseudonymizing Data
Thomas Vakili and Hercules Dalianis 318

Sentiment Classification of Historical Danish and Norwegian Literary Texts
Ali Allaith, Kirstine Nielsen Degn, Alexander Conroy, Bolette S. Pedersen, Jens Bjerring-Hansen and Daniel Hershcovich 324

Parser Evaluation for Analyzing Swedish 19th-20th Century Literature
Sara Stymne, Carin Östman and David Håkansson 335

An Empirical Study of Multitask Learning to Improve Open Domain Dialogue Systems
Mehrdad Farahani and Richard Johansson 347

Uncertainty-Aware Natural Language Inference with Stochastic Weight Averaging
Aarne Talman, Hande Celikkanat, Sami Virpioja, Markus Heinonen and Jörg Tiedemann 358

Alignment of Wikidata lexemes and Det Centrale Ordregister
Finn Årup Nielsen 366

Low-resource Bilingual Dialect Lexicon Induction with Large Language Models
Katya Artemova and Barbara Plank 371

Constructing a Knowledge Graph from Textual Descriptions of Software Vulnerabilities in the National Vulnerability Database
Anders Mølmen Høst, Pierre Lison and Leon Moonen 386

A Survey of Corpora for Germanic Low-Resource Languages and Dialects
Verena Blaschke, Hinrich Schuetze and Barbara Plank 392

You say tomato, I say the same: A large-scale study of linguistic accommodation in online communities
Aleksandrs Berdicevskis and Viktor Erbro 415

Integrating rules and neural nets for morphological tagging of Norwegian - Results and challenges
Dag Trygve Truslew Haug, Ahmet Yildirim, Kristin Hagen and Anders Nøklestad 425

Comparing Methods for Segmenting Elementary Discourse Units in a French Conversational Corpus
Laurent Prevot, Julie Hunter and Philippe Muller 436

Multi-way Variational NMT for UGC: Improving Robustness in Zero-shot Scenarios via Mixture Density Networks
José Carlos Rosales Núñez, Djamé Seddah and Guillaume Wisniewski 447

Multilingual Automatic Speech Recognition for Scandinavian Languages
Rafal Cerniavski and Sara Stymne 460

A character-based analysis of impacts of dialects on end-to-end Norwegian ASR
Phoebe Parsons, Knut Kvale, Torbjørn Svendsen and Giampiero Salvi 467

Quasi: a synthetic Question-Answering dataset in Swedish using GPT-3 and zero-shot learning
Dmytro Kalpakchi and Johan Boye 477

Automatic Closed Captioning for Estonian Live Broadcasts
Tanel Alumäe, Joonas Kalda, Külliki Bode and Martin Kaitsa 492

The Effect of Data Encoding on Relation Triplet Identification
Steinunn Rut Friðriksdóttir and Hafsteinn Einarsson 500

Improving Generalization of Norwegian ASR with Limited Linguistic Resources
Per Erik Solberg, Pablo Ortiz, Phoebe Parsons, Torbjørn Svendsen and Giampiero Salvi 508

The Finer They Get: Combining Fine-Tuned Models For Better Semantic Change Detection
Wei Zhou, Nina Tahmasebi and Haim Dubossarsky 518

Question Answering and Question Generation for Finnish
Ilmari Kylliäinen and Roman Yangarber 529

Probing structural constraints of negation in Pretrained Language Models
David Kletz, Marie Candito and Pascal Amsili 541

Boosting Norwegian Automatic Speech Recognition
Javier De La Rosa, Rolv-Arild Braaten, Per Egil Kummervold and Freddy Wetjen 555

Length Dependence of Vocabulary Richness
Niklas Zechner 565

A query engine for L1-L2 parallel dependency treebanks
Arianna Masciolini 574

Filtering Matters: Experiments in Filtering Training Sets for Machine Translation
Steinþór Steingrímsson, Hrafn Loftsson and Andy Way 588

Gamli - Icelandic Oral History Corpus: Design, Collection and Evaluation
Luke O'Brien, Finnur Ágúst Ingimundarson, Jón Guðnason and Steinþór Steingrímsson . . . 601

NoCoLA: The Norwegian Corpus of Linguistic Acceptability
Matias Jentoft and David Samuel 610

NorBench – A Benchmark for Norwegian Language Models
David Samuel, Andrey Kutuzov, Samia Touileb, Erik Velldal, Lilja Øvrelid, Egil Rønningstad, Elina Sigdel and Anna Sergeevna Palatkina 618

Making Instruction Finetuning Accessible to Non-English Languages: A Case Study on Swedish Models
Oskar Holmström and Ehsan Doostmohammadi 634

GiellaLT — a stable infrastructure for Nordic minority languages and beyond
Flammie A Pirinen, Sjur N. Moshagen and Katri Hiovain-Asikainen 643

Adapting an Icelandic morphological database to Faroese
Kristján Rúnarsson and Kristin Bjarnadottir 650

Danish Clinical Named Entity Recognition and Relation Extraction
Martin Sundahl Laursen, Jannik Skyttegaard Pedersen, Rasmus Søgaard Hansen, Thiusius Rajeeth Savarimuthu and Pernille Just Vinholt 655

Scaling-up the Resources for a Freely Available Swedish VADER (svVADER)
Dimitrios Kokkinakis, Ricardo Muñoz Sánchez and Mia-Marie Hammarlin 667

Colex2Lang: Language Embeddings from Semantic Typology
Yiyi Chen, Russa Biswas and Johannes Bjerva 673

Toxicity Detection in Finnish Using Machine Translation
Anni Eskelinen, Laura Silvala, Filip Ginter, Sampo Pyysalo and Veronika Laippala 685

Evaluating a Universal Dependencies Conversion Pipeline for Icelandic
Þórunn Arnardóttir, Hinrik Hafsteinsson, Atli Jasonarson, Anton Karl Ingaon and Steinþór Steingrímsson 698

Automatic Transcription for Estonian Children's Speech
Agnes Luhtaru, Rauno Jaaska, Karl Kruusamäe and Mark Fishel 705

Translated Benchmarks Can Be Misleading: the Case of Estonian Question Answering
Hele-Andra Kuulmets and Mark Fishel 710

Predicting the presence of inline citations in academic text using binary classification
Peter Vajdecka, Elena Callegari, Desara Xhura and Atli Snær Ásmundsson 717

Neural Text-to-Speech Synthesis for Võro
Liisa Rätsep and Mark Fishel 723

Transfer to a Low-Resource Language via Close Relatives: The Case Study on Faroese
Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš and Ivan Vulić 728

Evaluating Morphological Generalisation in Machine Translation by Distribution-Based Compositionality Assessment
Anssi Moisio, Mathias Creutz and Mikko Kurimo 738

Estonian Named Entity Recognition: New Datasets and Models
Kairit Sirts 752

Machine Translation for Low-resource Finno-Ugric Languages
Lisa Yankovskaya, Maali Tars, Andre Tättar and Mark Fishel 762

Distilling Estonian Text Domains for Production-Oriented Machine Translation
Elizaveta Korotkova and Mark Fishel 772

Spelling Correction for Estonian Learner Language
Kais Allkivi-Metsoja and Jaagup Kippar 782