

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

Tallinn, Estonia
2 March 2025

Editors:

**Špela Arhar Holdt
Nikolai Ilinykh
Barbara Scalvini**

**Micaella Bruton
Iben Nyholm Debess
Crina Madalina Tudor**

ISBN: 979-8-3313-1889-5

Printed from e-media with permission by:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571



Some format issues inherent in the e-media version may also appear in this print version.

Copyright© (2025) by Tartu Library, Estonia
All rights reserved.

Printed with permission by Curran Associates, Inc. (2025)

For permission requests, please contact the Association for Computational Linguistics
at the address below.

Association for Computational Linguistics
209 N. Eighth Street
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006
Fax: 1-570-476-0860

acl@aclweb.org

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Universal Dependencies Treebank for Uzbek</i> Arofat Akhundjanova and Luigi Talamo	1
<i>Fine-Tuning Cross-Lingual LLMs for POS Tagging in Code-Switched Contexts</i> Shayaan Absar	7
<i>Second language Korean Universal Dependency treebank v1.2: Focus on Data Augmentation and Annotation Scheme Refinement</i> Hakyung Sung and Gyu-Ho Shin	13
<i>Recommendations for Overcoming Linguistic Barriers in Healthcare: Challenges and Innovations in NLP for Haitian Creole</i> Ludovic Mompelat	20
<i>Beyond a Means to an End: A Case Study in Building Phonotactic Corpora for Central Australian Languages</i> Saliha Muradoglu, James Gray, Jane Helen Simpson, Michael Proctor and Mark Harvey	32
<i>OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches</i> Jenna Kanerva, Cassandra Ledins, Siiri Käpyaho and Filip Ginter	38
<i>FoQA: A Faroese Question-Answering Dataset</i> Annika Simonsen, Dan Saattrup Nielsen and Hafsteinn Einarsson	48
<i>Automatic Validation of the Non-Validated Spanish Speech Data of Common Voice 17.0</i> Carlos Daniel Hernández Mena, Barbara Scalvini and Dávid í Lág	58
<i>WikiQA-IS: Assisted Benchmark Generation and Automated Evaluation of Icelandic Cultural Knowledge in LLMs</i> Þórunn Arnardóttir, Elías Bjartur Einarsson, Garðar Ingvarsson Juto, Þorvaldur Páll Helgason and Hafsteinn Einarsson	64
<i>DUDU: A Treebank for Ottoman Turkish in UD Style</i> Enes Yılandiloğlu and Janine Siewert	74
<i>A Simple Audio and Text Collection-Annotation Tool Targeted to Brazilian Indigenous Language Native Speakers</i> Gustavo Padilha Polleti, Fabio Cozman and Fabricio Gerardi	80
<i>First Steps in Benchmarking Latvian in Large Language Models</i> Inguna Skadina, Bruno Bakanovs and Roberts Dargis	86
<i>On the Usage of Semantics, Syntax, and Morphology for Noun Classification in IsiZulu</i> Imaan Sayed, Zola Mahlaza, Alexander van der Leek, Jonathan Mopp and C. Maria Keet	96
<i>Annotating Attitude in Swedish Political Tweets</i> Anna Lindahl	106
<i>VerbCraft: Morphologically-Aware Armenian Text Generation Using LLMs in Low-Resource Settings</i> Hayastan Avetisyan and David Broneske	111
<i>Post-OCR Correction of Historical German Periodicals using LLMs</i> Vera Danilova and Gijs Aangenendt	120

<i>From Words to Action: A National Initiative to Overcome Data Scarcity for the Slovene LLM</i>	
Špela Arhar Holdt, Špela Antloga, Tina Munda, Eva Pori and Simon Krek	130
<i>Assessing the Similarity of Cross-Lingual Seq2Seq Sentence Embeddings Using Low-Resource Spectral Clustering</i>	
Nelson Moll and Tahseen Rabbani	137
<i>Voices of Luxembourg: Tackling Dialect Diversity in a Low-Resource Setting</i>	
Nina Hosseini-Kivanani, Christoph Schommer and Peter Gilles	143
<i>The Application of Corpus-Based Language Distance Measurement to the Diatopic Variation Study (on the Material of the Old Novgorodian Birchbark Letters)</i>	
Iliia Afanasev and Olga Lyashevskaya	153
<i>"I Need More Context and an English Translation": Analysing How LLMs Identify Personal Information in Komi, Polish, and English</i>	
Nikolai Ilinykh and Maria Irena Szawerna	165
<i>Multi-label Scandinavian Language Identification (SLIDE)</i>	
Mariia Fedorova, Jonas Sebulon Frydenberg, Victoria Handford, Victoria Ovedie Chruickshank Langø, Solveig Helene Willoch, Marthe Løken Midtgaard, Yves Scherrer, Petter Mæhlum and David Samuel	179
<i>Federated Meta-Learning for Low-Resource Translation of Kirundi</i>	
Kyle Rui Sang, Tahseen Rabbani and Tianyi Zhou	190