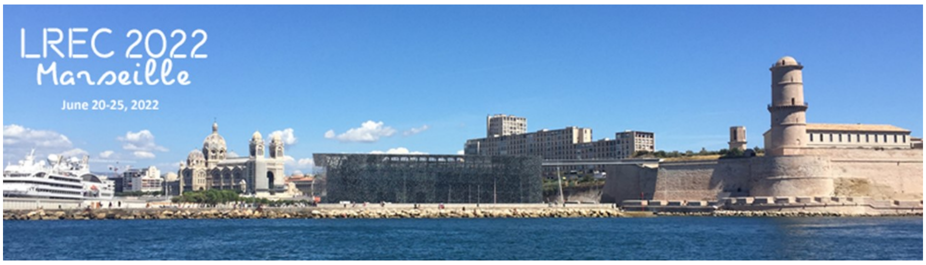

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

First Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL)

PROCEEDINGS

Editors: Jonne Sälevä, Constantine Lignos

Proceedings of the First Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL 2022)

Edited by: Jonne Sälevä and Constantine Lignos

ISBN: 978-2-493814-06-7
EAN: 9782493814067

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Message from the Organizers

This volume documents the Proceedings of the First Workshop on Dataset Creation for Lower-Resource Languages (DCLRL), held on June 24th, 2022 as part of the 13th Language Resources and Evaluation Conference (LREC 2022).

In recent years, there has been a significant increase in interest in developing datasets for lower-resourced languages (LRLs) and a greater involvement of the communities speaking those languages in the process. Developing resources for languages that have had fewer resources created for them poses a unique set of technical and ethical challenges that differs from higher-resourced language work.

The overall goal of this workshop was to create a new venue where previously disjoint research communities working on different areas surrounding lower-resourced languages can come together and share their insights across specialized research niches. We endorsed an open and intersectional perspective to the definition of a "lower-resourced language," acknowledging that this designation is both imperfect and often the result of many contributing factors. Our workshop was designed to be open and inclusive, presenting great scholarship from as many different perspectives as possible.

Papers submitted to the workshop were expected to generally revolve around resource creation for lower-resourced languages, but could be fairly broad in scope. For example, we welcomed submissions describing both finished or ongoing research projects, downloadable resources, and position papers containing insights on resource creation for lower-resourced languages that the broader community could benefit from.

The submissions that we received led us to slightly broaden the scope of the workshop to also welcome work in what might be termed *lower-resourced domains*; domains and tasks that are in need of more or higher quality datasets, even if these datasets are not necessarily created in languages that might be considered lower-resourced. We encourage organizers of future workshops and conferences to explicitly include this type of work in their calls for papers.

We are delighted to publish the ten papers that appear in the proceedings of our workshop, and we hope you will find them both informative and thought-provoking. We want to acknowledge that this workshop is of smaller scope than we had originally planned. Like many other workshops and conferences this year, our organizing process was affected by a number of external factors. Chief among them were the effects of the COVID-19 pandemic. Beyond the health and safety considerations, the pandemic has impacted the costs and logistics of conference travel and created additional workload and burnout in the research community.

We found ourselves carrying an exceptional workload from our university in this exceptional time, and therefore we opted to have only a minimal set of organizers and keep the workshop small and focused. We would like to acknowledge and thank the researchers with whom we had originally discussed the proposal for this workshop: David Adelani, Ximena Gutierrez-Vasques, Mmanape Hlungwane, Vukosi Marivate, and Priscilla Tyulu. We hope that in future iterations, we are able to engage an even broader portion of the community and increase the scale of this workshop.

Organizers

Constantine Lignos
Chester Palen-Michel
Jonne Sälevä

Program Committee

Linda Achilles
Petra Bago
Steven Bedrick
Stergios Chatzikyriakidis
Aparna Dutta
Hafsteinn Einarsson
Steinunn Rut Friðriksdóttir
Imane Guellil
Rejwanul Haque
Asha Hegde
Chaak-ming Lau
Jackson Lee
Muxuan Liu
Alex Lưu
Vukosi Marivate
Malte Ollmann
Hilary Prichard
Karthika Ranganathan
Caitlin Richter
Hosahalli Lakshmaiah Shashirekha
Ridouane Tachicart

Table of Contents