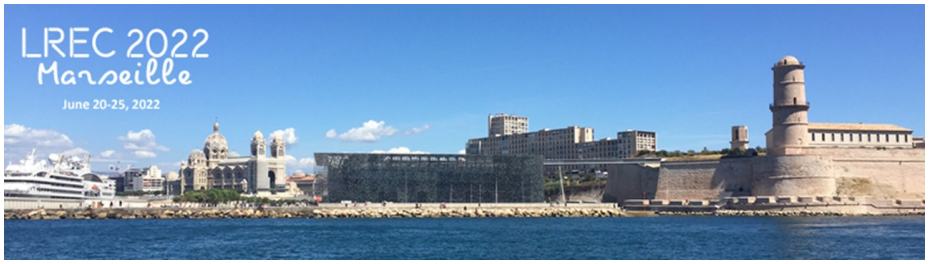

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

The 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL2022)

PROCEEDINGS

Editors:
Maite Melero
Sakriani Sakti
Claudia Soria

Sponsored by
Google

Proceedings of the LREC 2022 Workshop of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022)

Edited by:
Maite Melero, Sakriani Sakti, Claudia Soria

ISBN: 979-10-95546-91-7
EAN: 9791095546917

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 Workshop Chairs

Over the last years, research in text and speech processing for less-resourced languages has taken momentum. Initiatives and events have flourished, as well as hackathons, toolkits, special interest groups, and journals' special issues. The topic of less-resourced languages has ceased to be niche and has gained space in major conferences such as LREC, ACL, and Interspeech.

The multiplication of research interest makes it even more necessary for the community that revolves around less-resourced languages to find opportunities for aggregation and discussion. It is also very important that these occasions leave space for communities and representatives of under-resourced and endangered languages, in order to ensure that the research and development of technological solutions are in line with the needs and demands of those communities, with a view to open and inclusive research with strong social impact.

The 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages (SIGUL 2022) spans the research interest areas of less-resourced, under-resourced, endangered, minority and minoritized languages. SIGUL 2022 carries on the tradition of the CCURL-SLTU (Collaboration and Computing for Under-Resourced Languages – Spoken Language Technologies for Under-resourced languages) Workshop Series, which has been organized since 2008 and, as LREC Workshops, since 2014. As usual, SIGUL provides a forum for the presentation of cutting edge research in text and speech processing for under-resourced languages to both academic and industry researchers. In addition, it offers a venue where researchers in different disciplines and from varied backgrounds can fruitfully explore new areas of intellectual and practical development while honouring their common interest in sustaining less-resourced languages.

In order to promote synergies and to increase cross-fertilization between neighbouring disciplines, this year's workshop holds a joint session together with the 18th Workshop on Multiword Expressions (MWE 2022) and hosts a shared task on unsupervised Machine Translation techniques for the benefit of under-resourced languages, organized by the MT4All project (CEF 2019-EU-IA-0031).

This year, we have the pleasure to welcome 19 oral and 8 poster presentations, addressing a vast array of topics in NLP, Speech, Data and General issues. Accepted papers display a huge variety of languages, covering 76 different languages from Europe, Asia, Africa and the Americas. This workshop, together with at least five other LREC2022 workshops in neighbouring topics and the main conference track on less-resourced and endangered languages, clearly show how the topic of language resources and speech and natural language processing for less-resourced language is now a mature and well-established field.

The SIGUL 2022 workshop is organised and sponsored by the SIGUL organization, which serves as the Special Interest Group in under-resourced languages for both ELRA and ISCA associations. It is also endorsed by SIGEL, the ACL special interest group on endangered languages. In addition, this year's event has received a sponsorship grant from Google Inc.

Organizers

Maite Melero – Barcelona Supercomputing Center, Spain
Sakriani Sakti – JAIST, Japan
Claudia Soria – CNR-ILC, Italy

Program Committee:

Gilles Adda (LIMSI/IMMI CNRS, France)
Tunde Adegbola (African Language Technology Initiative)
Manex Agirrezabal (University of Copenhagen, Denmark)
Shyam S Agrawal (KIIT, India)
Begona Altuna (University of the Basque Country, Spain)
Raghuram Mandyam Annasamy (Google, US)
Antti Arppe (University of Alberta, Canada)
Dorothee Beermann (NTNU, Norway)
Delphine Bernhardt (Lilpa, Université de Strasbourg, France)
Laurent Besacier (Naver Labs Europe, France)
Steven Bird (Charles Darwin University, Australia)
Federico Boschetti (CNR-ILC, Italy)
Klara Ceberio Berger (Elhuyar, Spain)
Matt Coler (University of Groningen, Campus Fryslân, The Netherlands)
Omar Farooq (ZH College of Engineering and Technology, India)
Dafydd Gibbon (Bielefeld University, Germany)
Itziar Gonzalez-Dios (University of the Basque Country, Spain)
Jeff Good (University at Buffalo, USA)
Atticus Harrigan (University of Alberta, Canada)
Lars Hellan (NTNU, Norway)
Dewi Bryn Jones (Bangor University, UK)
John Judge (ADAPT DCU, Ireland)
Alexey Karpov (SPC RAS, Russian Federation)
Heysem Kaya (Utrecht University, The Netherlands)
Laurent Kevers (Università di Corsica Pasquale Paoli, France)
Irina Kipyatkova (SPC RAS, Russian Federation)
Andras Kornai (Hungarian Academy of Sciences, Hungary)
Jordan Lachler (University of Alberta, Canada)
Richard Littauer (University of Saarland, Germany)
Joseph Mariani (LIMSI-CNRS, France)
Satoshi Nakamura (NAIST, Japan)
Win Pa Pa (UCS Yangon, Myanmar)
Delyth Prys (Bangor University, UK)
Carlos Ramisch (Université Marseille, France)
Kevin Scannell (Saint Louis University, Missouri, US)
Nick Thieberger (University of Melbourne / ARC Centre of Excellence for the Dynamics of Language, Australia)
Trond Trosterud (Tromsø University, Norway)
Daan Van Esch (Google)
Charl Van Heerden (Saigen (Pty) Ltd, South Africa)
Marcely Zanon Boito (LIA – Avignon University, France)

Table of Contents

Workshop Program

Friday, June 24, 2022

14:00–14:10 SIGUL 2022 Opening Talk

14:10–15:10 Session 1: Speech

14:10–14:25 *Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings*
Marcely Zanon Boito, Bolaji Yusuf, Lucas Ondel, Aline Villavicencio and Laurent Besacier

14:25–14:40 *An Open Source Web Reader for Under-Resourced Languages*
Judy Fong, Þorsteinn Daði Gunnarsson, Sunneva Þorsteinsdóttir, Gunnar Thor Örnólfsson and Jon Gudnason

14:40–14:55 *Text-to-Speech for Under-Resourced Languages: Phoneme Mapping and Source Language Selection in Transfer Learning*
Phat Do, Matt Coler, Jelske Dijkstra and Esther Klabbers

14:55–15:10 *ReadAlong Studio: Practical Zero-Shot Text-Speech Alignment for Indigenous Language Audiobooks*
Patrick Littell, Eric Joanis, Aidan Pine, Marc Tessier, David Huggins Daines and Delasie Torkornoo

15:10–16:00 Keynote Speech
Sovereignty for Under-resourced Languages
Keoni Mahelona

16:00–16:30 Coffee break

16:30–17:45 Session 2: Data

16:30–16:45 *Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text*
Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha and Bharathi Raja Chakravarthi

16:45–17:00 *Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish*
Mathilde Hutin and Marc Allassonnière-Tang

17:00–17:15 *Tupían Language Ressources: Data, Tools, Analyses*
Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon and Fabrício F. Gerardi

Friday, June 24, 2022 (continued)

17:15–17:30 *Quality versus Quantity: Building Catalan-English MT Resources*
Ona de Gibert Bonet, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé and Maite Melero

17:30–17:45 *A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context*
Ronny Mabokela and Tim Schlippe

Saturday, June 25, 2022

9:00–10:00 Session 3: MT4All

9:00–9:15 *General overview of unsupervised MT for under resourced languages*
Â Jordi Armengol

9:15–9:30 *Technical approach in MT4All*
Iakes Goenaga

9:30–9:45 *MT4All generated resources and Shared Task scope and results*
Ona de Gibert

9:45–10:00 *CUNI Submission to MT4All Shared Task*
Ivana Kvapilíková and Ondrej Bojar

10:00–10:30 Session 4: General Issues

10:00–10:15 *Resource: Indicators on the Presence of Languages in Internet*
Daniel Pimienta

10:15–10:30 *Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights*
A. Seza Doğruöz and Sunayana Sitaram

10:30–11:00 Coffee break

11:00–12:45 Session 5: NLP

11:00–11:15 *Sentiment Analysis for Hausa: Classifying Students' Comments*
Ochilbek Rakhmanov and Tim Schlippe

11:15–11:30 *Nepali Encoder Transformers: An Analysis of Auto Encoding Transformer Language Models for Nepali Text Classification*
Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel and Bal Krishna Bal

11:30–11:45 *CoSwID, a Code Switching Identification Method Suitable for Under-Resourced Languages*
Laurent Kevers

Saturday, June 25, 2022 (continued)

11:45–12:00 *A Neural Network Approach to Create Minangkabau-Indonesia Bilingual Dictionary*
Kartika Resiandi, Yohei Murakami and Arbi Haza Nasution

12:00–12:15 *Machine Translation from Standard German to Alemannic Dialects*
Louisa Lambrecht, Felix Schneider and Alexander Waibel

12:15–12:30 *Question Answering Classification for Amharic Social Media Community Based Questions*
Tadesse Destaw, Seid Muhie Yimam, Abinew Ayele and Chris Biemann

12:30–12:45 *Automatic Detection of Morphological Processes in the Yorùbá Language*
Tunde Adegbola

12:45–14:00 Lunch break

14:00–15:00 Joint SIGUL2022-MWE Poster session

Evaluating Unsupervised Approaches to Morphological Segmentation for Wolastoqey
Diego Bear and Paul Cook

Baseline English and Maltese-English Classification Models for Subjectivity Detection, Sentiment Analysis, Emotion Analysis, Sarcasm Detection, and Irony Detection
Keith Cortis and Brian Davis

Building Open-source Speech Technology for Low-resource Minority Languages with SáMi as an Example – Tools, Methods and Experiments
Katri Hiovain-Asikainen and Sjur Moshagen

Investigating the Quality of Static Anchor Embeddings from Transformers for Under-Resourced Languages
Pranaydeep Singh, Orphee De Clercq and Els Lefever

Introducing YakuToolkit. Yakut Treebank and Morphological Analyzer.
Tatiana Merzhevich and Fabrício Ferraz Gerardi

A Language Model for Spell Checking of Educational Texts in Kurdish (Sorani)
Roshna Abdulrahman and Hossein Hassani

SimRelUz: Similarity and Relatedness Scores as a Semantic Evaluation Dataset for Uzbek Language
Ulugbek Salaev, Elmurod Kuriyozov and Carlos Gómez-Rodríguez

ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot
Dimitra Anastasiou

Saturday, June 25, 2022 (continued)

15:00–16:00 **Joint SIGUL2022-MWE Keynote Speech**
Multiword Expressions and the Low-Resource Scenario from the Perspective of a Local Oral Culture
Steven Bird

16:00–16:30 **Coffee break**

16:30–17:30 **Panel discussion**

17:30–17:50 **General discussion**

17:50–18:00 **Closing**