

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

Held at LREC 2022

Marseille, France  
20 – 25 June 2022

## **Editors:**

**Maite Melero  
Sakriani Sakti  
Claudia Soria**

ISBN: 978-1-7138-6138-6

**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© (2022) by European Language Resources Association (ELRA)  
All rights reserved.

Copyright for individual papers remains with the authors and are licensed under a Creative Commons 4.0 license, CC-BY-NC. (<https://creativecommons.org/licenses/by-nc/4.0/>)

Printed with permission by Curran Associates, Inc. (2023)

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](mailto: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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

<i>Unsupervised Word Segmentation from Discrete Speech Units in Low-Resource Settings</i> Marcely Zanon Boito, Bolaji Yusuf, Lucas Ondel, Aline Villavicencio and Laurent Besacier . . . . .	1
<i>An Open Source Web Reader for Under-Resourced Languages</i> Judy Fong, Þorsteinn Daði Gunnarsson, Sunneva Þorsteinsdóttir, Gunnar Thor Örnólfsson and Jon Guðnason . . . . .	10
<i>Text-to-Speech for Under-Resourced Languages: Phoneme Mapping and Source Language Selection in Transfer Learning</i> Phat Do, Matt Coler, Jelske Dijkstra and Esther Klabbers . . . . .	16
<i>ReadAlong Studio: Practical Zero-Shot Text-Speech Alignment for Indigenous Language Audiobooks</i> Patrick Littell, Eric Joanis, Aidan Pine, Marc Tessier, David Huggins Daines and Delasie Torkornoo . . . . .	23
<i>Corpus Creation for Sentiment Analysis in Code-Mixed Tulu Text</i> Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, Hosahalli Lakshmaiah Shashirekha and Bharathi Raja Chakravarthi . . . . .	33
<i>Crowd-sourcing for Less-resourced Languages: Lingua Libre for Polish</i> Mathilde Hutin and Marc Allasonnière-Tang . . . . .	41
<i>Tupían Language Ressources: Data, Tools, Analyses</i> Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon and Fabrício F. Gerardi . . . . .	48
<i>Quality versus Quantity: Building Catalan-English MT Resources</i> Ona de Gibert Bonet, Ksenia Kharitonova, Blanca Calvo Figueras, Jordi Armengol-Estapé and Maite Melero . . . . .	59
<i>A Sentiment Corpus for South African Under-Resourced Languages in a Multilingual Context</i> Ronny Mabokela and Tim Schlippe . . . . .	70
<i>CUNI Submission to MT4All Shared Task</i> Ivana Kvapilíková and Ondrej Bojar . . . . .	78
<i>Resource: Indicators on the Presence of Languages in Internet</i> Daniel Pimienta . . . . .	83
<i>Language Technologies for Low Resource Languages: Sociolinguistic and Multilingual Insights</i> A. Seza Dođruöz and Sunayana Sitaram . . . . .	92
<i>Sentiment Analysis for Hausa: Classifying Students' Comments</i> Ochilbek Rakhmanov and Tim Schlippe . . . . .	98
<i>Nepali Encoder Transformers: An Analysis of Auto Encoding Transformer Language Models for Nepali Text Classification</i> Utsav Maskey, Manish Bhatta, Shiva Bhatt, Sanket Dhungel and Bal Krishna Bal . . . . .	106
<i>CoSwID, a Code Switching Identification Method Suitable for Under-Resourced Languages</i> Laurent Kevers . . . . .	112

<i>A Neural Network Approach to Create Minangkabau-Indonesia Bilingual Dictionary</i> Kartika Resiandi, Yohei Murakami and Arbi Haza Nasution .....	122
<i>Machine Translation from Standard German to Alemannic Dialects</i> Louisa Lambrecht, Felix Schneider and Alexander Waibel .....	129
<i>Question Answering Classification for Amharic Social Media Community Based Questions</i> Tadesse Destaw, Seid Muhie Yimam, Abinew Ayele and Chris Biemann .....	137
<i>Automatic Detection of Morphological Processes in the Yorùbá Language</i> Tunde Adegbola .....	146
<i>Evaluating Unsupervised Approaches to Morphological Segmentation for Wolastoqey</i> Diego Bear and Paul Cook .....	155
<i>Baseline English and Maltese-English Classification Models for Subjectivity Detection, Sentiment Analysis, Emotion Analysis, Sarcasm Detection, and Irony Detection</i> Keith Cortis and Brian Davis .....	161
<i>Building Open-source Speech Technology for Low-resource Minority Languages with SáMi as an Example – Tools, Methods and Experiments</i> Katri Hiovain-Asikainen and Sjur Moshagen .....	169
<i>Investigating the Quality of Static Anchor Embeddings from Transformers for Under-Resourced Languages</i> Pranaydeep Singh, Orphee De Clercq and Els Lefever .....	176
<i>Introducing YakuToolkit. Yakut Treebank and Morphological Analyzer.</i> Tatiana Merzhevich and Fabrício Ferraz Gerardi .....	185
<i>A Language Model for Spell Checking of Educational Texts in Kurdish (Sorani)</i> Roshna Abdulrahman and Hossein Hassani .....	189
<i>SimRelUz: Similarity and Relatedness Scores as a Semantic Evaluation Dataset for Uzbek Language</i> Ulugbek Salaev, Elmurod Kuriyozov and Carlos Gómez-Rodríguez .....	199
<i>ENRICH4ALL: A First Luxembourgish BERT Model for a Multilingual Chatbot</i> Dimitra Anastasiou .....	207