

# **First Workshop on Dataset Creation for Lower-Resourced Languages (DCLRL 2022)**

Held at LREC 2022

Marseille, France  
20 – 25 June 2022

## **Editors:**

**Jonne Saleva**  
**Constantine Lignos**

ISBN: 978-1-7138-6117-1

**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 the Association for Computational Linguistics  
All rights reserved.

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

Printed with permission by Curran Associates, Inc. (2022)

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>SyntAct: A Synthesized Database of Basic Emotions</i> Felix Burkhardt, Florian Eyben and Björn Schuller .....	1
<i>Data Sets of Eating Disorders by Categorizing Reddit and Tumblr Posts: A Multilingual Comparative Study Based on Empirical Findings of Texts and Images</i> Christina Baskal, Amelie Elisabeth Beutel, Jessika Keberlein, Malte Ollmann, Esra Üresin, Jana Vischinski, Janina Weihe, Linda Achilles and Christa Womser-Hacker .....	10
<i>Construction and Validation of a Japanese Honorific Corpus Based on Systemic Functional Linguistics</i> Muxuan Liu and Ichiro Kobayashi .....	19
<i>Building an Icelandic Entity Linking Corpus</i> Steinunn Rut Friðriksdóttir, Valdimar Ágúst Eggertsson, Benedikt Geir Jóhannesson, Hjalti Daníels-son, Hrafn Loftsson and Hafsteinn Einarsson .....	27
<i>Crawling Under-Resourced Languages - A Portal for Community-Contributed Corpus Collection</i> Erik Körner, Felix Helfer, Christopher Schröder, Thomas Eckart and Dirk Goldhahn .....	36
<i>Fine-grained Entailment: Resources for Greek NLI and Precise Entailment</i> Eirini Amanaki, Jean-Philippe Bernardy, Stergios Chatzikyriakidis, Robin Cooper, Simon Dobnik, Aram Karimi, Adam Ek, Eirini Chrysovalantou Giannikouri, Vasiliki Katsouli, Ilias Kolokousis, Eirini Chrysovalantou Mamatzaki, Dimitrios Papadakis, Olga Petrova, Erofilii Psaltaki, Charikleia Soupiona, Effrosyni Skoulataki and Christina Stefanidou .....	44
<i>Words.hk: a Comprehensive Cantonese Dictionary Dataset with Definitions, Translations and Translit-erated Examples</i> Chaak-ming Lau, Grace Wing-yan Chan, Raymond Ka-wai Tse and Lilian Suet-ying Chan .....	53
<i>LiSTra Automatic Speech Translation: English to Lingala Case Study</i> Salomon Kabongo Kabenamualu, Vukosi Marivate and Herman Kamper .....	63
<i>Ara-Women-Hate: An Annotated Corpus Dedicated to Hate Speech Detection Against Women in the Arabic Community</i> Imane Guellil, Ahsan Adeel, Faical Azouaou, Mohamed Boubred, Yousra Houichi and Akram Abdelhaq Moumna .....	68
<i>Word-level Language Identification Using Subword Embeddings for Code-mixed Bangla-English Social Media Data</i> Aparna Dutta .....	76