

10th Workshop on Challenges in the Management of Large Corpora (CMLC-10)

Held at LREC 2022

Marseille, France
20 – 25 June 2022

Editors:

**Piotr Banski
Adrien Barbaresi
Simon Clemenide**

**Marc Kupietz
Harald Lungen**

ISBN: 978-1-7138-6115-7

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

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>Challenges in Creating a Representative Corpus of Romanian Micro-Blogging Text</i> Vasile Pais, Maria Mitrofan, Verginica Barbu Mititelu, Elena Irimia, Roxana Micu and Carol Luca Gasan	1
<i>Exhaustive Indexing of PubMed Records with Medical Subject Headings</i> Modest von Korff	8
<i>UDEasy: a Tool for Querying Treebanks in CoNLL-U Format</i> Luca Brigada Villa	16
<i>Matrix and Double-Array Representations for Efficient Finite State Tokenization</i> Nils Diewald	20
<i>Count-Based and Predictive Language Models for Exploring DeReKo</i> Peter Fankhauser and Marc Kupietz	27
<i>“The word expired when that world awoke.” New Challenges for Research with Large Text Corpora and Corpus-Based Discourse Studies in Totalitarian Times</i> Hanno Biber	32