

8th Workshop on Challenges in the Management of Large Corpora (CMLC-8)

Marseille, France
11 – 16 May 2020

Editors:

**Piotr Banski
Adrien Barbaresi
Simon Clemenide**

**Marc Kupietz
Harald Lungen
Ines Pisetta**

ISBN: 978-1-7138-1257-9

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© (2020) 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. (2020)

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>Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora</i> Denis Arnold, Bernhard Fisseni, Pawel Kamocki, Oliver Schonefeld, Marc Kupietz and Thomas Schmidt	1
<i>Evaluating a Dependency Parser on DeReKo</i> Peter Fankhauser, Bich-Ngoc Do and Marc Kupietz	10
<i>French Contextualized Word-Embeddings with a sip of CaBeRnet: a New French Balanced Reference Corpus</i> Murielle Popa-Fabre, Pedro Javier Ortiz Suárez, Benoît Sagot and Éric de la Clergerie	15
<i>Geoparsing the historical Gazetteers of Scotland: accurately computing location in mass digitised texts</i> Rosa Filgueira, Claire Grover, Melissa Terras and Beatrice Alex	24
<i>The Corpus Query Middleware of Tomorrow – A Proposal for a Hybrid Corpus Query Architecture</i> Markus Gärtner	31
<i>Using full text indices for querying spoken language data</i> Elena Frick and Thomas Schmidt	40
<i>Challenges for Making Use of a Large Text Corpus such as the ‘AAC – Austrian Academy Corpus’ for Digital Literary Studies</i> Hanno Biber	47
<i>Czech National Corpus in 2020: Recent Developments and Future Outlook</i> Michal Kren	52
<i>Adding a Syntactic Annotation Level to the Corpus of Contemporary Romanian Language</i> Andrei Scutelnicu, Catalina Maranduc and Dan Cristea	58