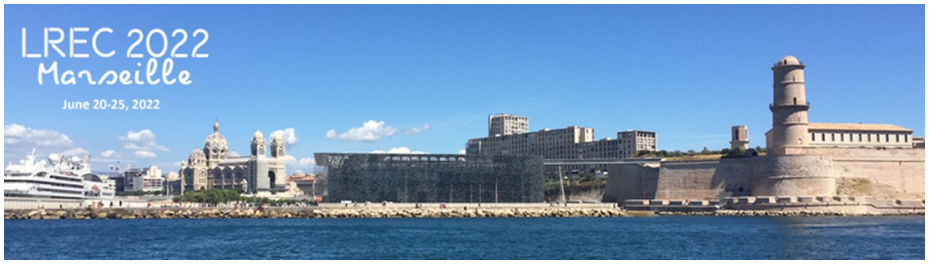

LREC 2022 Workshop
Language Resources and Evaluation Conference
20-25 June 2022

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

PROCEEDINGS

Editors:
Piotr Bański, Adrien Barbaresi, Simon Clematide,
Marc Kupietz, Harald Lüngen

Proceedings of the LREC 2022 Workshop on Challenges in the Management of Large Corpora (CMLC-10 2022)

Edited by:
Piotr Bański, Adrien Barbaresi, Simon Clematide, Marc Kupietz, Harald Lüngen

ISBN: 979-10-95546-83-2
EAN: 9791095546832

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Preface

Creating very large corpora no longer appears to be a challenge. With the constantly growing amount of born-digital text – be it available on the web or only on the servers of publishing companies – and with the rising number of printed texts digitised by public institutions or technological giants such as Google, we may safely expect the upper limits of text collections to keep increasing for years to come. Although some of this was already true 20 years ago, we have a strong impression that the challenge has now shifted from an increase in terms of size to the effective and efficient processing of the large amounts of primary data and much larger amounts of annotation data.

On the one hand, some fundamental technical methods and strategies call for re-evaluation. These include, for example, efficient and sustainable curation of data, management of collections that span multiple volumes or that are distributed across several centres, innovative corpus architectures that maximise the usefulness of data, and techniques that allow for efficient search and analysis.

On the other hand, the new challenges require research into language-modelling methods and new corpus-linguistic methodologies that can make use of extremely large, semi-structured datasets. These methodologies must re-address the tasks of investigating rare phenomena involving multiple lexical items, of finding and representing fine-grained sub-regularities, and of investigating variations within and across language domains. This should be accompanied by new methods to structure both content and search results, in order to, among others, cope with false positives, assess data quality, or ensure interoperability. Another much-needed research goal is visualisation techniques that facilitate the interpretation of results and formulation of new hypotheses.

Due to the interest that the first meeting of CMLC (held at LREC-2012 in Istanbul) enjoyed, the workshop became a cyclic event. The second meeting took place at LREC again, in 2014 in Reykjavík; the third edition of CMLC was part of Corpus Linguistics 2015 in Lancaster. The fourth meeting took place in Portorož, Slovenia, as part of LREC-2016. CMLC-5 was an event combined with BigNLP-2017 and took place as part of the Corpus Linguistics conference in Birmingham. The sixth meeting took us to Japan (LREC-2018 in Miyazaki), and the seventh to Wales (CL 2019 in Cardiff). Due to the COVID-19 pandemic, the eighth event, scheduled to be co-located with LREC-2020 in Marseille, shared the fate of the conference and was cancelled at the post-review stage, while we chose to maintain the event numbering for the sake of the proceedings volume. The subsequent meeting, at CL 2021, organised by the University of Limerick, was fully virtual.

In 2022, as part of this year's LREC, we are going to meet in hybrid mode, the physical part of which is going to be Marseille. The leading questions for papers and discussions during CMLC-10 are: (a) What can be done to deal with IPR and data protection issues? (b) What sampling techniques can we apply? (c) What quality issues should we be aware of? (d) What infrastructures and frameworks are being developed for the efficient storage, annotation, analysis and retrieval of large datasets? (e) What affordances do visualisation techniques offer for the exploratory analysis approaches of corpora? (f) What kinds of APIs or other means of access would make the corpus data as widely usable as possible without interfering with legal restrictions? (g) How to guarantee that corpus data remain available and sustainably usable?

We would like to thank the Authors and the Programme Committee for their effort, and we are looking forward to meeting or seeing many of you in person, at last.

Piotr Bański, Adrien Barbaresi, Simon Clematide, Marc Kupietz, Harald Lüngen

Organizers

Piotr Bański – Leibniz-Institut für Deutsche Sprache, Mannheim
Adrien Barbaresi – Berlin-Brandenburg Academy of Sciences
Simon Clematide – University of Zurich
Marc Kupietz – Leibniz-Institut für Deutsche Sprache, Mannheim
Harald Lüngen – Leibniz-Institut für Deutsche Sprache, Mannheim

Program Committee:

Laurence Anthony, Waseda University (Japan)
Vladimír Benko, Slovak Academy of Sciences (Slovakia)
Damir Ćavar, Indiana University (USA)
Nils Diewald, IDS Mannheim (Germany)
Tomaž Erjavec, Jožef Stefan Institute, Ljubljana (Slovenia)
Johannes Graën, University of Zurich (Switzerland)
Andrew Hardie, Lancaster University (UK)
Serge Heiden, ENS de Lyon/IHRIM (France)
Miloš Jakubíček, Lexical Computing Ltd. (UK)
Paweł Kamocki, IDS Mannheim (Germany)
Natalia Kotsyba, Samsung (Poland)
Dawn Knight, Cardiff University (UK)
Michal Křen, Charles University, Prague (Czech Republic)
Veronika Laippala, University of Turku (Finland)
Verena Lyding, EURAC Research (Italy)
Paul Rayson, Lancaster University (UK)
Laurent Romary, INRIA (France)
Jan-Oliver Rüdiger, IDS Mannheim (Germany)
Roman Schneider, IDS Mannheim (Germany)
Serge Sharoff, University of Leeds (UK)
Irena Spasić, Cardiff University (UK)
Marko Tadić, University of Zagreb (Croatia)
Ludovic Tanguy, University of Toulouse (France)
Tamás Váradi, Hungarian Academy of Sciences (Hungary)
Andreas Witt, IDS / University of Mannheim (Germany)

Table of Contents

Workshop Program

Monday, June 20, 2022

09:00–10:30 Session 1

9:00–9:15 *Technical Setup and Welcome*

9:15–9:30 *Intro*

9:30–10:00 *Challenges in Creating a Representative Corpus of Romanian Micro-Blogging Text*
Vasile Pais, Maria Mitrofan, Verginica Barbu Mititelu, Elena Irimia, Roxana Micu and Carol Luca Gasan

10:00–10:30 *Exhaustive Indexing of PubMed Records with Medical Subject Headings*
Modest von Korff

10:00–11:00 Coffee Break

11:00–13:00 Session 2

11:00–11:30 *UDeasy: a Tool for Querying Treebanks in CoNLL-U Format*
Luca Brigada Villa

11:30–12:00 *Matrix and Double-Array Representations for Efficient Finite State Tokenization*
Nils Diewald

12:00–12:30 *Count-Based and Predictive Language Models for Exploring DeReKo*
Peter Fankhauser and Marc Kupietz

12:30–13:00 *"The word expired when that world awoke." New Challenges for Research with Large Text Corpora and Corpus-Based Discourse Studies in Totalitarian Times*
Hanno Biber