LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

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

PROCEEDINGS

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

Proceedings of the LREC 2020
8th Workshop on Challenges in the Management of Large Corpora (CMLC-8)

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

ISBN: 979-10-95546-61-0
EAN: 9791095546610

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

Introduction

In order to satisfy the information needs of a wide range of researchers across a number of disciplines, large textual datasets require careful design, collection, cleaning, encoding, annotation, storage, retrieval, and curation. This daunting set of tasks has coalesced into a number of key themes and questions that are of interest to the contributing research communities: (a) what sampling techniques can we apply? (b) what quality issues should we be aware of? (c) what infrastructures and frameworks are being developed for the efficient storage, annotation, analysis and retrieval of large datasets? (d) what affordances do visualisation techniques offer for the exploratory analysis approaches of corpora? (e) what legal paths can be followed in dealing with IPR and data protection issues governing both the data sources and the query results? (f) how to guarantee that corpus data remain available and usable in a sustainable way?

Over the past years, the CMLC workshop series has gathered researchers interested in these long-standing topics while also willing to address newly developing trends in the field. At each meeting, we also made sure to reserve space for national corpus project reports. In the year 2020, we expected our meeting to be co-located with the LREC conference, which had unfailingly hosted the previous CMLC events, from Istanbul through Reykjavík, Portorož and Miyazaki. Unfortunately, due to the spreading COVID-19 pandemic, the May date had to be cancelled. Nevertheless, we are hereby offering the present volume of proceedings, with thanks to the contributing Authors and heartfelt gratitude to the Reviewers. Whether we will be able to suggest an alternative date or mode for the meeting, time will tell. The relevant information will be placed on the CMLC-8 homepage at http://corpora.ids-mannheim.de/cmlc-2020.html .

P. Bański, A. Barbaresi, S. Clematide, M. Kupietz, H. Lüngen May 2020

Organizers:

Piotr Bański, IDS Mannheim
Adrien Barbaresi, Berlin-Brandenburg Academy of Sciences
Simon Clematide, Institute of Computational Linguistics, UZH
Marc Kupietz, IDS Mannheim
Harald Lüngen, IDS Mannheim

Program Committee:

Laurence Anthony, Waseda University, Japan
Vladimír Benko, Slovak Academy of Sciences
Felix Bildhauer, IDS Mannheim
Sonja Bosch, University of South Africa
Dan Cristea, "Alexandru Ioan Cuza" University of Iasi
Damir Ćavar, Indiana University, Bloomington
Tomaž Erjavec, Jožef Stefan Institute
Stefan Evert, Friedrich-Alexander-Universität Nürnberg/Erlangen
Johannes Graën, University of Gothenburg, Pompeu Fabra University
Andrew Hardie, Lancaster University
Serge Heiden, ENS de Lyon
Miloš Jakubíček, Lexical Computing Ltd.
Dawn Knight, Cardiff University, UK
Natalia Kotsyba, Samsung Poland
Michal Křen, Charles University, Prague
Sandra Kübler, Indiana University, Bloomington
Gaël Lejeune, Sorbonne Université
Paul Rayson, Lancaster University
Martin Reynaert, Tilburg University
Laurent Romary, INRIA
Kevin Scannell, Saint-Louis University
Roland Schäfer, FU Berlin
Serge Sharoff, University of Leeds
Irena Spasic, Cardiff University
Marko Tadić, University of Zagreb, Faculty of Humanities and Social Sciences
Ludovic Tanguy, University of Toulouse
Dan Tufiş, Romanian Academy, Bucharest
Tamás Váradi, Research Institute for Linguistics, Hungarian Academy of Sciences

Table of Contents

Conference Program

Addressing Cha(lle)nges in Long-Term Archiving of Large Corpora
Denis Arnold, Bernhard Fisseni, Pawel Kamocki, Oliver Schonefeld, Marc Kupietz and Thomas Schmidt

Evaluating a Dependency Parser on DeReKo
Peter Fankhauser, Bich-Ngoc Do and Marc Kupietz

French Contextualized Word-Embeddings with a sip of CaBeRnet: a New French Balanced Reference Corpus
Murielle Popa-Fabre, Pedro Javier Ortiz Suárez, Benoît Sagot and Éric de la Clergerie

Geoparsing the historical Gazetteers of Scotland: accurately computing location in mass digitised texts
Rosa Filgueira, Claire Grover, Melissa Terras and Beatrice Alex

The Corpus Query Middleware of Tomorrow – A Proposal for a Hybrid Corpus Query Architecture
Markus Gärtner

Using full text indices for querying spoken language data
Elena Frick and Thomas Schmidt

Challenges for Making Use of a Large Text Corpus such as the 'AAC – Austrian Academy Corpus' for Digital Literary Studies
Hanno Biber

Czech National Corpus in 2020: Recent Developments and Future Outlook
Michal Kren

Adding a Syntactic Annotation Level to the Corpus of Contemporary Romanian Language
Andrei Scutelnicu, Catalina Maranduc and Dan Cristea