LREC 2026

The 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026

Workshop Proceedings

Editors:

Piotr Bański
Dawn Knight
Marc Kupietz
Andreas Witt
Alina Wróblewska

May 11, 2026

Proceedings of the 12th Workshop on
Challenges in the Management of Large Corpora (CMLC-12)

ISBN 978-2-493814-67-8

Preface

The twelfth meeting of the workshop on the Challenges in the Management of Large Corpora has brought us back to where it all started: a welcoming embrace of the LREC conference series. As in the previous CMLC meetings, we have decided to explore common areas of interest across a range of issues in language resource management, corpus linguistics, natural language processing, natural language generation, and data science.

Large textual datasets require careful design, collection, cleaning, encoding, annotation, storage, retrieval, and curation to be of use for a wide range of research questions and to users across a number of disciplines. A growing number of national and other very large corpora are being made available, many historical archives are being digitised, numerous publishing houses are opening their textual assets for text mining, and many billions of words can be quickly sourced from the web and online social media.

A mixed blessing of our times is that much of those texts, in mono- and multi-lingual arrangements, can now be created automatically by exploiting Large Language Models at various scales. That, on the one hand, makes it possible to inflate the amounts of data where normally data would be scarce: in under-resourced languages or language varieties, in specific genres or for intricate and rarely attested constructions. On the other hand, such procedures immediately raise concerns regarding the authenticity and quality of such data, casting doubt on the possibility of adequately (truthfully, verifiably, reproducibly) addressing the kind of research questions that prompted the rapid but tainted increase of the available data volumes in the first place. Similar doubts may be directed at mass creation of secondary and tertiary data ordinarily crucial for linguistic research: apart from potential legal constraints on the use of the original body of human-created data, new questions arise as to the legal status of the derived data, the ways to create (among others) provenance metadata of the derived resources, and the level of trust regarding mass-produced grammatical (and other) annotation layers.

These new as well as more traditional questions lie at the base of the list of topics that management of large corpora (for any currently suitable definition of "large") invokes or at least strongly brushes against.

CMLC has often invited reports on broadly understood national corpus initiatives. Given that it has been a while since the last round, we have decided to host a "What's the news?" session, with some of our veteran presenters as well as colleagues who have not yet introduced their national corpus projects. The scale of the response to our invitation has been a welcome surprise and we are happy to devote part of this volume to (broadly defined) national corpus projects. This is intended as a snapshot of the current state, with a proud look at the history (including developments since some of the projects were last presented at previous CMLCs), and a somewhat wary glance at what the future may bring.

P. Bański, D. Knight, M. Kupietz, A. Witt, A. Wróblewska
April 2026

Organising Committee

Piotr Bański (IDS Mannheim)
Dawn Knight (Cardiff University)
Marc Kupietz (IDS Mannheim)
Andreas Witt (IDS Mannheim)
Alina Wróblewska (ICS PAS, Warsaw)

Programme Committee

Laurence Anthony (Waseda University, Japan)
Vladimír Benko (Slovak Academy of Sciences)
Felix Bildhauer (IDS Mannheim)
Mark Davies (English-Corpora.org)
Nils Diewald (IDS Mannheim)
Kaja Dobrovoljc (University of Ljubljana / Jožef Stefan Institute)
Jarle Ebeling (University of Oslo)
Tomaž Erjavec (Jožef Stefan Institute, Ljubljana)
Andrew Hardie (Lancaster University, UK)
Serge Heiden (ENS de Lyon)
Ulrich Heid (University of Hildesheim)
Nancy Ide (Vassar College / Brandeis University)
Olha Kanishcheva (Heidelberg University)
Gražina Korvel (Vilnius University)
Natalia Kocyba (Samsung Poland)
Michal Křen (Charles University, Prague)
Anna Latusek (ICS PAS, Warsaw)
Paul Rayson (Lancaster University)
Laurent Romary (INRIA)
Thomas Schmidt (University of Duisburg-Essen)
Serge Sharoff (University of Leeds)
Maria Shvedova (Kharkiv Polytechnic Institute / University of Jena)
Irena Spasić (Cardiff University)
Martin Wynne (University of Oxford)

Table of Contents

Section I: Managing Large Corpora

Section II: News from National Corpus Initiatives

Workshop Programme

9:00–9:10 *Welcome and introduction*
Piotr Bański

9:10–9:40 **Session A: Short papers**
Chair: Dawn Knight

9:10–9:25 *TestiMole-Conversational: A 30-Billion-Word Italian Discussion Board Corpus (1996–2024) for Language Modeling and Sociolinguistic Research*
Matteo Rinaldi, Rossella Varvara and Viviana Patti

9:25–9:40 *IfGPT, a Large Dataset Representing Bulgarian, with the Bulgarian National Corpus as Its Core*
Svetla Peneva Koeva and Ivelina Stoyanova

9:45–10:25 **Session B: Flash presentations of posters**
Chairs: Alina Wróblewska, Piotr Bański

Hellenic National Corpus: The Current State
Maria Gavriilidou and Nikolaos Sidiropoulos

Corpas Náisiúnta Na Gaeilge 2022-2029: A Project Overview
Mícheál J. Ó Meachair, Úna Bhreathnach, Kevin Scannell, Michal Mechura, Brian Ó Raghallaigh and Gearóid Ó Cleircín

General Regionally Annotated Corpus of Ukrainian: Recent Developments and Future Plans
Maria Shvedova

Recent Developments of the Bulgarian National Corpus
Svetla Peneva Koeva and Ivelina Stoyanova

The British National Corpus 1994 to 2026
Martin Wynne and Megan Bushnell

The Corpus of Contemporary Polish: 2011-2020 Decade and Beyond
Witold Kieraś, Małgorzata Marciniak, Katarzyna Krasnowska-Kieraś and Marcin Woliński

Building the v4 of the Croatian National Corpus
Marko Tadić, Vanja Štefanec and Daša Farkaš

Managing Growth in a National Corpus: The Hungarian National Corpus 3.0 (MNSZ3)
Noémi Ligeti-Nagy, Enikő Héja, Ágnes Bánfi, Flóra Földesi, Bence Sárossy, Boglárka Skrabák, Tamás Váradi and Gábor Prószéky

CoRoLa Version 2.0: Corpus Enrichment and a New Annotation Level
Elena Irimia, Verginica Barbu Mititelu, Radu Ion, Vasile Pais, Maria Mitrofan and Dan Ioan Tufis

The German Medical Text Corpus: Early 2026 Update
Justin Hofenbitzer, Christina Lohr, Frank Meineke, Markus Löffler and Martin Boeker

From Corpus to Community: New NLP Tools for Welsh Language Research and Learning
Dawn Knight and Fernando Alva-Manchego

Swiss-AL: Language Data Platform for Applied Sciences
Julia Krasselt, Philipp Dreesen, Dolores Lemmenmeier-Batinić, Sooyeon Geckeler, Klaus Rothenhäusler and Matthias Fluor

EuReCo, KorAP and DeReKo: Updates on Ingestion and Annotation Pipelines, Backend, Interfaces, Operation, and Corpora
Marc Kupietz, Nils Diewald, Harald Lüngen, Eliza Margaretha Illig, Helge Stallkamp, Uyen-Nhu Tran and Rameela Yaddehige

Merimënga: A Manifest-First Pipeline for Reproducible Albanian Web Corpus Construction
Besim Kabashi and Michael Ruppert

10:40–11:25 **Session 3: Poster session**
Chair: Andreas Witt

11:30–13:00 **Session 4: Long papers**
Chair: Marc Kupietz

11:30–12:00 *Pop Lyrics through Time: Challenges in Corpus-Based Modeling of Linguistic and Emotional Dynamics in German Pop Lyrics*
Roman Schneider

12:00–12:30 *The Infrastructure behind Latvian National Corpora Collection*
Roberts Dargis and Baiba Valkovska

12:30–13:00 *Optimized for AI: Curating the Icelandic Gigaword Corpus for Stable LLM Training*
Jón Friðrik Daðason and Steinþór Steingrímsson