COLING 2025

18th Workshop on Building and Using Comparable Corpora

PROCEEDINGS

20 January, 2025

©Copyright The International Committee on Computational Linguistics (ICCL), 2025
These proceedings are licensed under a Creative Commons
Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 979-8-89176-211-4
ISSN 2951-2093 (COLING)

Message from the Program Chairs

The 18th Workshop on Building and Using Comparable Corpora
(BUCC) @ COLING 2025

In the language engineering and linguistics communities, research in comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is chiefly motivated by the need to use comparable corpora as training data for statistical NLP applications such as statistical and neural machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest because they enable cross-language discoveries and comparisons. It is generally accepted in both communities that comparable corpora consist of documents that are comparable in content and form in various degrees and dimensions across several languages or language varieties. Parallel corpora are on the one end of this spectrum, unrelated corpora on the other.

Comparable corpora have been used in various applications, including Information Retrieval, Machine Translation, Cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for statistical natural language processing applications, for example, to extract parallel corpora from comparable corpora for neural machine translation. As such, it is of great interest to bring together builders and users of such corpora. The aim of the workshop series on "Building and Using Comparable Corpora" (BUCC) is to promote progress in this field.

The previous editions of the workshop took place in Africa (LREC 2008 in Marrakech), America (ACL 2011 in Portland and ACL 2017 in Vancouver), Asia (ACL-IJCNLP 2009 in Singapore, ACL-IJCNLP 2015 in Beijing, LREC 2018 in Miyazaki, Japan), Europe (LREC 2010 in Malta, ACL 2013 in Sofia, LREC 2014 in Reykjavik, LREC 2016 in Portoroz, RANLP 2019 and RANLP 2023 in Varna, LREC 2022 in Marseille, LREC-COLING-2024 in Turin) and also on the border between Asia and Europe (LREC 2012 in Istanbul). Due to the Corona crisis, the workshop was also held online in conjunction with LREC 2020 and RANLP 2021. The materials of the past workshops and related studies have also been summarised in a recent textbook from Springer:
`https://link.springer.com/book/10.1007/978-3-031-31384-4`.

We want to thank all the people who, in one way or another, helped make this workshop once again a success, especially the COLING workshop chairs, and publication chairs.

Our special thanks go to our invited speakers, Ken Church and Preslav Nakov, and to the members of the program committee, who did a great job in reviewing the submitted papers under strict time constraints. Last but not least, we would like to thank the authors and all workshop participants.

Serge Sharoff, Ayla Rigouts Terryn, Pierre Zweigenbaum, Reinhard Rapp January 2025

Organizing Committee

Serge Sharoff University of Leeds, United Kingdom

Ayla Rigouts Terryn Université de Montréal, Mila, Canada

Pierre Zweigenbaum Université Paris-Saclay, CNRS, LISN, Orsay, France

Reinhard Rapp University of Mainz, Germany

Program Committee

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Eleftherios Avramidis (DFKI, Germany)
- Gabriel Bernier-Colborne (National Research Council, Canada)
- Thierry Etchegoyhen (Vicomtech, Spain)
- Natalia Grabar (University of Lille, France)
- Amal Haddad Haddad (Universidad de Granada, Spain)
- Kyo Kageura (University of Tokyo, Japan)
- Natalie Kübler (Université Paris Cité, France)
- Philippe Langlais (Université de Montréal, Canada)
- Yves Lepage (Waseda University, Japan)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)
- Dragos Stefan Munteanu (Language Weaver, Inc., USA)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Nasredine Semmar (CEA LIST, Paris, France)
- Silvia Severini (Leonardo Labs, Italy)
- Richard Sproat (OGI School of Science & Technology, USA)
- Tim Van de Cruys (KU Leuven, Belgium)
- François Yvon (Sorbonne Université, France)

Table of Contents

Conference Program

9:15–9:30 **Opening and introduction**

9:30–10:30 **Multilingual corpus development**

Bilingual resources for Moroccan Sign Language Generation and Standard Arabic Skills Improvement of Deaf Children
Abdelhadi Soudi, Corinne Vinopol and Kristof Van Laerhoven

Harmonizing Annotation of Turkic Postverbial Constructions: A Comparative Study of UD Treebanks
Arofat Akhundjanova

10:30–11:00 **Coffee break, morning**

11:00–13:00 **Multilinguality of Large Language Models**

Towards Truly Open, Language-Specific, Safe, Factual, and Specialized Large Language Models
Preslav Nakov

Make Satire Boring Again: Reducing Stylistic Bias of Satirical Corpus by Utilizing Generative LLMs
Asli Umay Ozturk, Recep Firat Cekinel and Pinar Karagoz

BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language
Ehsan Lotfi, Nikolay Banar and Walter Daelemans

13:00–14:00 Lunch

14:00–15:30 Machine Translation and Cross-lingual Processing

Refining Dimensions for Improving Clustering-based Cross-lingual Topic Models
Chia-Hsuan Chang, Tien Yuan Huang, Yi-Hang Tsai, Chia-Ming Chang and San-Yih Hwang

The Role of Handling Attributive Nouns in Improving Chinese-To-English Machine Translation
Adam Meyers, Rodolfo Joel Zevallos, John E. Ortega and Lisa Wang

Can a Neural Model Guide Fieldwork? A Case Study on Morphological Data Collection
Aso Mahmudi, Borja Herce, Demian Inostroza Améstica, Andreas Scherbakov, Eduard H. Hovy and Ekaterina Vylomova

15:30–16:00 Coffee break, afternoon

16:00–17:30 Diversity of language resources

Comparable Corpora: Opportunities for New Research Directions
Kenneth Ward Church

SELEXINI – a large and diverse automatically parsed corpus of French
Manon Scholivet, Agata Savary, Louis Estève, Marie Candito and Carlos Ramisch

17:30–17:45 Closing remarks