19th Workshop on Building and Using Comparable Corpora (BUCC)

Co-located with LREC 2026,
Palma de Mallorca,
May 11, 12, or 16, 2026 (exact day to be announced)
Submission deadline: 28 Feb 2026

MOTIVATION

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 data-driven 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. Parallel corpora are on the one end of this spectrum, and unrelated corpora are on the other. Increasingly, these resources are not only collected, but also augmented or even created synthetically, which raises new questions about how to define and measure comparability.

In recent years, the use of comparable corpora for pre-training Large Language Models (LLMs) has led to their impressive multilingual and cross-lingual abilities, which are relevant to a range of applications, including information retrieval, machine translation, cross-lingual text classification, etc. The linguistic definitions and observations related to comparable corpora are crucial to improve methods to mine such corpora, to assess and document synthetic data, and to improve cross-lingual transfer of LLMs. Therefore, it is of great interest to bring together builders and users of such corpora.

TOPICS

We solicit contributions on all topics related to comparable (and parallel) corpora, including but not limited to the following:

Building Comparable Corpora:

- Automatic and semi-automatic methods, including generating comparable corpora using LLMs
- Methods to mine parallel and non-parallel corpora from the web
- Tools and criteria to evaluate the comparability of corpora
- Parallel vs non-parallel corpora, monolingual corpora
- Rare and minority languages, within and across language families
- Multi-media/multi-modal comparable corpora

Synthetic Data for Comparable Corpora

- LLM generation of comparable/parallel data
- Improving comparability of synthetic data

- Incidental bilingualism & pre-training use of comparable data
- Comparability & cross-lingual consistency
- Detection & attribution of synthetic vs. human text
- English-centric effects & fairness across languages/scripts
- Evaluation & reproducibility for downstream tasks

Applications of comparable corpora:

- Human translation
- Language learning
- Cross-language information retrieval & document categorization
- Bilingual and multilingual projections
- (Unsupervised) machine translation
- Writing assistance
- Machine learning techniques using comparable corpora

Mining from Comparable Corpora:

- Cross-language distributional semantics, word embeddings and pre-trained multilingual transformer models
- Extraction of parallel segments or paraphrases from comparable corpora
- Methods to derive parallel from non-parallel corpora (e.g. to provide for low-resource languages in neural machine translation)
- Extraction of bilingual and multilingual translations of single words, multi-word expressions, proper names, named entities, sentences, paraphrases etc. from comparable corpora
- Induction of morphological, grammatical, and translation rules from comparable corpora
- Induction of multilingual word classes from comparable corpora

Comparable Corpora in the Humanities:

- Comparing linguistic phenomena across languages in contrastive linguistics
- Analyzing properties of translated language in translation studies
- Studying language change over time in diachronic linguistics
- Assigning texts to authors via authors' corpora in forensic linguistics
- Comparing rhetorical features in discourse analysis
- Studying cultural differences in sociolinguistics
- Analyzing language universals in typological research

Panel Discussion

The panel discusses the impact of synthetic data on comparable corpora research. Fundamental questions about how LLMs transform our understanding and use of multilingual data are addressed.

PRACTICAL INFORMATION

The workshop is a hybrid event, both in-person and online. Workshop registration is via the main conference registratio

The workshop proceedings will be published in the ACL Anthology.

IMPORTANT DATES

Deadlines are “anywhere on Earth.”

28 Feb 2026	Paper submission deadline
22 Mar 2026	Notification of acceptance
29 Mar 2026	Camera-ready final papers
14 Apr 2026	Workshop Programme final version
May 11, 12, or 16, 2026	Workshop date (exact day TBA)

For updates of the schedule, please follow the present Web page.

SUBMISSION GUIDELINES

Please follow the style sheet and templates (for LaTeX, Overleaf, and MS-Word) provided for the main conference.

Papers should be submitted as a PDF file using the START conference manager.

Submissions must describe original and unpublished work and range from 4 to 8 pages plus unlimited references. Reviewing will be double blind, so the papers should not reveal the authors’ identity. Accepted papers will be published in the workshop proceedings.

Double submission policy: Parallel submission to other meetings or publications is possible but must be notified to the workshop organizers by e-mail immediately upon submission to another venue.

For further information and updates see the present Web page.

INFORMATION ABOUT THE LRE 2026 MAP AND THE "SHARE YOUR LRs!" INITIATIVE

When submitting a paper from the START page, authors will be asked to provide essential information about language resources (LRs in a broad sense, i.e. also technologies, standards, evaluation kits, etc.) that have been used for the work described in the paper or are a new result of the research.

Moreover, ELRA encourages all LREC authors to share the described LRs (data, tools, services, etc.) to enable their reuse and replicability of experiments (including evaluation ones).

PDF CFP : bucc2026-cfp.pdf
Last modified: 10 Dec 2026

ORGANIZERS AND CONTACT

Reinhard Rapp (University of Mainz, Germany)

Ayla Rigouts Terryn (Université de Montréal (UdeM), Mila, Canada)

Serge Sharoff (University of Leeds, United Kingdom)

Pierre Zweigenbaum (Université Paris-Saclay, CNRS, France)

Contact: reinhardrapp (at) gmx (dot) de

PROGRAMME COMMITTEE

- Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
- Eleftherios Avramidis (DFKI, Germany)
- Gabriel Bernier-Colborne (National Research Council, Canada)
- Kenneth Church (VecML.com, USA)
- Patrick Drouin (Université de Montréal, Canada)
- Alex Fraser (Technical University of Munich, Germany)
- Natalia Grabar (CNRS, 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)
- Shervin Malmasi (Amazon, USA)
- Michael Mohler (Language Computer Corporation, USA)
- Emmanuel Morin (Nantes Université, France)
- Dragos Stefan Munteanu (RWS, USA)
- Preslav Nakov (Mohamed bin Zayed University of AI (MBZUAI), United Arab Emirates)
- Ted Pedersen (University of Minnesota, Duluth, USA)
- Reinhard Rapp (University of Mainz, Germany)
- Ayla Rigouts Terryn (Université de Montréal & Mila, Canada)
- Nasredine Semmar (CEA-List, France)
- Serge Sharoff (University of Leeds, UK)
- Richard Sproat (Sakana.ai, Tokyo, Japan)
- Marko Tadić (University of Zagreb, Croatia)
- François Yvon (CNRS & Sorbonne Université, France)
- Pierre Zweigenbaum (Université Paris-Saclay, CNRS, France)

Last modified: 10 Dec 2026