



LREC 2026
Palma de Mallorca, Spain

**The 19th Workshop on
Building and Using Comparable Corpora (BUCC)
@ LREC 2026**

Workshop Proceedings

**Editors
Reinhard Rapp, Ayla Rigouts Terryn,
Serge Sharoff, Pierre Zweigenbaum**

11 May 2026

Proceedings of the 19th Workshop on Building and Using Comparable Corpora (BUCC) @
LREC 2026

©ELRA Language Resources Association (ELRA), 2026
These proceedings are licensed under a Creative Commons Attribution-
NonCommercial 4.0 International License (CC BY-NC 4.0)

ISBN 978-2-493814-75-3

Preface

The 19th Workshop on Building and Using Comparable Corpora (BUCC) @ LREC 2026

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. 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, COLING 2025 in Abu Dhabi), 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 Torino), 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 LREC workshop chairs.

Our special thanks go to our invited speaker, Els Lefever, 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.

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

May 2026

Organizing Committee

Workshop Chairs

Reinhard Rapp (University of Mainz, Germany)

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

Serge Sharoff (University of Leeds, United Kingdom)

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

Program 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, United Kingdom)
- 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)

Table of Contents

<i>Keynote: The Cross-Lingual Transfer Myth: Why Modern LLMs Still Fail Without Comparable Corpora and Representations</i>	
Els Lefever	1
<i>A Comparative Study of Parkinsonian Speech Corpora for Deep Learning-Based Detection of Dysarthria</i>	
Clara Ponchard and Pierre Serrano	2
<i>Computing Semantic Similarity for Aligning Bilingual Semi-parallel Texts: A Case Study</i>	
Steffen Frenzel, Maximilian Krupop and Manfred Stede	9
<i>A Comparative Study in Corpus Linguistics Applied to Automatic Terminology Extraction</i>	
Mercè Vázquez, Sergi Alvarez-Vidal and Antoni Oliver	20
<i>Comparable Corpora in Cross-linguistic Research: Nominal Number in English, Czech, and Greek</i>	
Konstantinos Diamantopoulos and Magda Ševčíková	30
<i>Liebe Kolleg:innen, Querid@s Compañer@s: Presenting the GILDEES Corpus</i>	
Marie-Pauline Krielke	41
<i>A Diachronic Comparable Corpus of Spanish Digital News (2017–2026) for the Study of Stylistic Convergence in the GenAI Era</i>	
Hugo Sanjurjo-González	53
<i>Align and Shine: Building High-quality Sentence-aligned Corpora for Multilingual Text Simplification</i>	
Luis Kenji Hilaraca Sanchez, Nouran Khallaf and Serge Sharoff	62
<i>Bi-Text Mining across German Dialects: On the Role of Synthetic Training Data for Dialect Adaptation</i>	
Jing Wang, Barbara Plank and Robert Litschko	72
<i>Parallel Corpora of Scholarly Documents for English-French Machine Translation</i>	
Ziqian Peng, Lichao Zhu, Rachel Bawden, Maud Bénard, Éric de la Clergerie, Mathilde Huguin, Natalie Kübler, Paul Lerner, Alexandra Mestivier and François Yvon	84
<i>Validating a Pipeline to Create a Comparable Corpus of Government-Issued Travel Advisories from the Internet Archives</i>	
Laura Braun and Christian Oswald	96
<i>Leveraging Comparable Toxicity Lexicons in Prompt Instructions for Multilingual Text Detoxification</i>	
Yassir El Attar, Esra Dönmez, Nina K. Ohlendorf and Agnieszka Falenska	108

Workshop Program

Monday, May 11, 2026

9:00

Session 1

Chair: Ayla Rigouts Terryn

9:00

Introduction

9:06

Keynote: The Cross-Lingual Transfer Myth: Why Modern LLMs Still Fail Without Comparable Corpora and Representations

Els Lefever

10:06

A Comparative Study of Parkinsonian Speech Corpora for Deep Learning-Based Detection of Dysarthria

Clara Ponchard and Pierre Serrano

10:30

Coffee break

11:00

Session 2: Comparable corpora for linguistics research

Chair: Philippe Langlais

11:00

Computing Semantic Similarity for Aligning Bilingual Semi-parallel Texts: A Case Study

Steffen Frenzel, Maximilian Krupop and Manfred Stede

11:24

A Comparative Study in Corpus Linguistics Applied to Automatic Terminology Extraction

Mercè Vázquez, Sergi Alvarez-Vidal and Antoni Oliver

11:48

Comparable Corpora in Cross-linguistic Research: Nominal Number in English, Czech, and Greek

Konstantinos Diamantopoulos and Magda Ševčíková

12:12

Liebe Kolleg:innen, Querid@s Compañer@s: Presenting the GILDEES Corpus

Marie-Pauline Krielke

12:36

A Diachronic Comparable Corpus of Spanish Digital News (2017–2026) for the Study of Stylistic Convergence in the GenAI Era

Hugo Sanjurjo-González

Monday, May 11, 2026 (continued)

13:00 Lunch break

14:00 Session 3: Synthetic corpora
Chair: Serge Sharoff

14:00 *Panel discussion: How comparable are synthetic data?*

15:12 *Align and Shine: Building High-quality Sentence-aligned Corpora for Multi-lingual Text Simplification*
Luis Kenji Hilaraca Sanchez, Nouran Khallaf and Serge Sharoff

15:36 *Bi-Text Mining across German Dialects: On the Role of Synthetic Training Data for Dialect Adaptation*
Jing Wang, Barbara Plank and Robert Litschko

16:00 Coffee break

16:30 Session 4: Building comparable datasets
Chair: Pierre Zweigenbaum

16:30 *Parallel Corpora of Scholarly Documents for English-French Machine Translation*
Ziqian Peng, Lichao Zhu, Rachel Bawden, Maud Bénard, Éric de la Clergerie, Mathilde Huguin, Natalie Kübler, Paul Lerner, Alexandra Mestivier and François Yvon

16:54 *Validating a Pipeline to Create a Comparable Corpus of Government-Issued Travel Advisories from the Internet Archives*
Laura Braun and Christian Oswald

17:18 *Leveraging Comparable Toxicity Lexicons in Prompt Instructions for Multilingual Text Detoxification*
Yassir El Attar, Esra Dönmez, Nina K. Ohlendorf and Agnieszka Falenska

17:42 Closing words