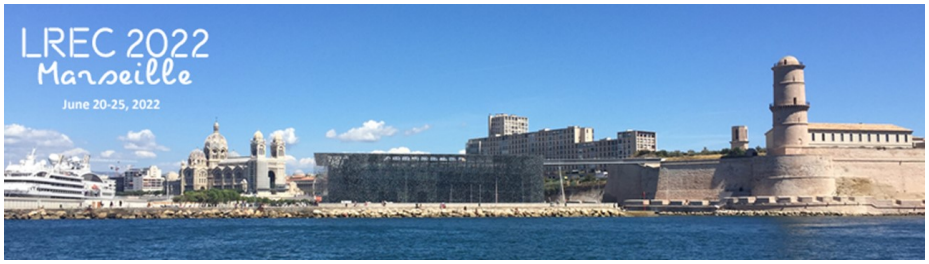

LREC 2022 Workshop
Language Resources and Evaluation Conference
25 June 2022

15th Workshop on Building and Using Comparable Corpora (BUCC 2022)

PROCEEDINGS

Editors:
Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

Proceedings of the LREC 2022
15th Workshop on Building and Using Comparable Corpora (BUCC 2022)

Edited by:
Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

ISBN: 979-10-95546-94-8
EAN: 9791095546948

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

Preface – 15th BUCC at LREC 2022)

This volume documents the Proceedings of the 15th Workshop on Building and Using Comparable Corpora, held on June 25, 2022, as part of the LREC 2022 conference (International Conference on Language Resources and Evaluation).

In the language engineering and the linguistics communities, research on comparable corpora has been motivated by two main reasons. In language engineering, on the one hand, it is primarily motivated by the need to use comparable corpora as training data for statistical Natural Language Processing applications such as statistical machine translation or cross-lingual retrieval. In linguistics, on the other hand, comparable corpora are of interest in themselves by making possible inter-linguistic discoveries and comparisons. It is generally accepted in both communities that comparable corpora are documents in one or several languages that are comparable in content and form in various degrees and dimensions. We believe that the linguistic definitions and observations related to comparable corpora can improve methods to mine such corpora for applications of statistical NLP. As such, it is of great interest to bring together builders and users of such corpora.

Comparable corpora are collections of documents that are comparable in content and form in various degrees and dimensions. This definition includes many types of parallel and non-parallel multilingual corpora, but also sets of monolingual corpora that are used for comparative purposes. Research on comparable corpora is active but used to be scattered among many workshops and conferences. The workshop series on "Building and Using Comparable Corpora" (BUCC) aims at promoting progress in this exciting field by bundling its research, thereby making it more visible and giving it a better platform.

The first 12 of the 14 previous editions of the workshop took place in Africa (LREC'08 in Marrakech), America (ACL'11 in Portland and ACL'17 in Vancouver), Asia (ACL-IJCNLP'09 in Singapore, ACL-IJCNLP'15 in Beijing, LREC'18 in Miyazaki, Japan), Europe (LREC'10 in Malta, ACL'13 in Sofia, LREC'14 in Reykjavik, LREC'16 in Portoroz, RANLP'19 in Varna) and also on the border between Asia and Europe (LREC'12 in Istanbul). Due to the Corona crises, in the past two years the conference was held online in conjunction with LREC'20 and with RANLP'21.

Part of this year's edition of the BUCC workshop was a shared task on "Bilingual Term Alignment in Comparable Specialized Corpora" which is documented in these proceedings..

We would like to thank all people who in one way or another helped in making this workshop once again a success. We are especially grateful to Khalid Choukri for his extraordinary guidance concerning the proceedings, to Nicoletta Calzolari for her continuous support of our workshop, and to Hélène Mazo, Sara Goggi and the whole team of LREC organisers for finding solutions to all matters of concern.

Our special thanks go to our invited speakers and to the members of the programme committee who did an excellent job in reviewing the submitted papers under strict time constraints. Last but not least we would like to thank our authors, shared task teams and all participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff June 2022

Workshop Organizers

Reinhard Rapp, Athena R.C., Magdeburg-Stendal University of Applied Sciences, University of Mainz (workshop chair)
Pierre Zweigenbaum, Université Paris-Saclay, CNRS, LISN (shared task chair)
Serge Sharoff, University of Leeds

Programme Committee

Ahmet Aker (University of Sheffield, UK)
Ebrahim Ansari (Institute for Advanced Studies in Basic Sciences, Iran)
Thierry Etchegoyhen (VicomTech, Spain)
Hitoshi Isahara (Otemon Gakuin University, Japan)
Kyo Kageura (The University of Tokyo, Japan)
Natalie Kübler (Université de Paris, France)
Philippe Langlais (Université de Montréal, Canada)
Yves Lepage (Waseda University, Japan)
Michael Mohler (Language Computer Corp., USA)
Emmanuel Morin (Université de Nantes, France)
Dragos Stefan Munteanu (Language Weaver, Inc., USA)
Ted Pedersen (University of Minnesota, Duluth, USA)
Reinhard Rapp (Athena R.C., Greece, Magdeburg-Stendal University of Applied Sciences, University of Mainz, Germany)
Nasredine Semmar (CEA LIST, Paris, France)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (OGI School of Science & Technology, USA)
Pierre Zweigenbaum (Université Paris-Saclay, CNRS, LISN, Orsay, France)

Table of Contents

Evaluating Monolingual and Crosslingual Embeddings on Datasets of Word Association Norms
Trina Kwong, Emmanuele Chersoni and Rong Xiang 1

About Evaluating Bilingual Lexicon Induction
Martin Laville, Emmanuel Morin and Phillippe Langlais 8

Don't Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings
Silvia Severini, Viktor Hangya, Masoud Jalili Sabet, Alexander Fraser and Hinrich Schütze 15

Building Domain-specific Corpora from the Web: the Case of European Digital Service Infrastructures
Rik van Noord, Cristian García-Romero, Miquel Esplà-Gomis, Leopoldo Pla Sempere and Antonio Toral 23

Multilingual Comparative Analysis of Deep-Learning Dependency Parsing Results Using Parallel Corpora
Diego Alves, Marko Tadić and Božo Bekavac 33

CUNI Submission to the BUCC 2022 Shared Task on Bilingual Term Alignment
Borek Požár, Klára Tauchmanová, Kristýna Neumannová, Ivana Kvapilíková and Ondřej Bojar . 43

Challenges of Building Domain-Specific Parallel Corpora from Public Administration Documents
Filip Klubička, Lorena Kasunić, Danijel Blazsetin and Petra Bago 50

Setting Up Bilingual Comparable Corpora with Non-Contemporary Languages
Helena Bermudez Sabel, Francesca Dell'Oro, Cyrielle Montrichard and Corinne Rossari 56

Fusion of linguistic, neural and sentence-transformer features for improved term alignment
Andraz Repar, Senja Pollak, Matej Ulčar and Boshko Koloski 61

Workshop Programme

09:00–9:05 ***Opening***

Session 1: Invited Presentation

09:05–10:00

Session 2: Comparative dependency parsing

10:00–10:30 *Multilingual Comparative Analysis of Deep-Learning Dependency Parsing Results Using Parallel Corpora*
Diego Alves, Marko Tadić and Božo Bekavac

10:30–11:00 ***Coffee Break***

Session 3: Building corpora and lexicon induction

11:00–11:30 *Building Domain-specific Corpora from the Web: the Case of European Digital Service Infrastructures*
Rik van Noord, Cristian García-Romero, Miquel Esplà-Gomis, Leopoldo Pla Sempere and Antonio Toral

11:30–12:00 *Challenges of Building Domain-Specific Parallel Corpora from Public Administration Documents*
Filip Klubička, Lorena Kasunić, Danijel Blazsetin and Petra Bago

12:00–12:30 *Setting Up Bilingual Comparable Corpora with Non-Contemporary Languages*
Helena Bermudez Sabel, Francesca Dell'Oro, Cyrielle Montrichard and Corinne Rossari

12:30–13:00 *About Evaluating Bilingual Lexicon Induction*
Martin Laville, Emmanuel Morin and Phillippe Langlais

13:00–14:00 ***Lunch Break***

Session 4: Word embeddings

14:00–14:30 *Evaluating Monolingual and Crosslingual Embeddings on Datasets of Word Association Norms*
Trina Kwong, Emmanuele Chersoni and Rong Xiang

14:30–15:00 *Don't Forget Cheap Training Signals Before Building Unsupervised Bilingual Word Embeddings*
Silvia Severini, Viktor Hangya, Masoud Jalili Sabet, Alexander Fraser and Hinrich Schütze

Session 5: Shared task on bilingual term alignment

15:00–15:30 *CUNI Submission to the BUCC 2022 Shared Task on Bilingual Term Alignment*
Borek Požár, Klára Tauchmanová, Kristýna Neumannová, Ivana Kvapilíková and Ondřej Bojar

15:30–16:00 *Fusion of linguistic, neural and sentence-transformer features for improved term alignment*
Andraz Repar, Senja Pollak, Matej Ulčar and Boshko Koloski

Workshop Programme (continued)

16:00–16:30 ***Coffee Break***

16:30–17:00 *Overview on the shared task*
Pierre Zweigenbaum

17:00–17:10 ***Closing***