

# Ninth Workshop on Building and Using Comparable Corpora

## Workshop Programme

Monday, May 23, 2016

**09.15–9.25** *Opening Remarks*

**Session 1: Invited Presentation**

09.25–10.30 Ruslan Mitkov

*The Name of the Game is Comparable Corpora*

**10.30–11.00** *Coffee Break*

**Session 2: Building Comparable Corpora**

11:00–11:30 Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko and Lyubov Ivanova

*Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints*

11:30–12:00 Yong Xu and François Yvon

*A 2D CRF Model for Sentence Alignment*

12:00–12:30 Mehdi Mohammadi

*Parallel Document Identification using Zipf's Law*

**12.30–14.00** *Lunch Break*

**Session 3: Invited Presentation**

14.00–15.00 Gregory Grefenstette

*Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research*

**Session 4: Applications of Comparable Corpora**

15.00–15.30 Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang and Shizhuang Weng

*A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction*

10:00–10:30 Ana Sabina Uban

*Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation*

**16.00–16.30** *Coffee Break*

**Session 5: Discussion**

16:30–17:30 Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp

*Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora*

**17.30–17.35** *Closing*



## **Editors**

Reinhard Rapp, University of Mainz, Germany

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

Serge Sharoff, University of Leeds, UK

## **Workshop Programme Committee**

Ahmet Aker, University of Sheffield (UK)

Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)

Éric Gaussier (Université Joseph Fourier, Grenoble, France)

Vishal Goyal (Punjabi University, Patiala, India)

Gregory Grefenstette (INRIA, Saclay, France)

Silvia Hansen-Schirra (University of Mainz, Germany)

Hitoshi Isahara (Toyohashi University of Technology)

Kyo Kageura (University of Tokyo, Japan)

Philippe Langlais (Université de Montréal, Canada)

Shervin Malmasi (Harvard Medical School, Boston, MA, USA)

Michael Mohler (Language Computer Corp., US)

Emmanuel Morin (Université de Nantes, France)

Dragos Stefan Munteanu (Language Weaver, Inc., US)

Lene Offersgaard (University of Copenhagen, Denmark)

Ted Pedersen (University of Minnesota, Duluth, US)

Reinhard Rapp (University of Mainz, Germany)

Serge Sharoff (University of Leeds, UK)

Michel Simard (National Research Council Canada)

Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)

## **Invited Speakers**

Gregory Grefenstette, INRIA Saclay, Université Paris Saclay, France)

Ruslan Mitkov, University of Wolverhampton, UK



# Table of Contents

<i>The Name of the Game is Comparable Corpora</i> Ruslan Mitkov .....	1
<i>Clustering Comparable Corpora of Russian and Ukrainian Academic Texts: Word Embeddings and Semantic Fingerprints</i> Andrey Kutuzov, Mikhail Kopotev, Tatyana Sviridenko and Lyubov Ivanova .....	3
<i>A 2D CRF Model for Sentence Alignment</i> Yong Xu and François Yvon .....	11
<i>Parallel Document Identification using Zipf’s Law</i> Mehdi Mohammadi .....	21
<i>Exploring the Richness and Limitations of Web Sources for Comparable Corpus Research</i> Gregory Grefenstette .....	26
<i>A Mutual Iterative Enhancement Model for Simultaneous Comparable Corpora and Bilingual Lexicons Construction</i> Zede Zhu, Xinhua Zeng, Shouguo Zheng, Xiongwei Sun, Shaoqi Wang and Shizhuang Weng .	27
<i>Hard Synonymy and Applications in Automatic Detection of Synonyms and Machine Translation</i> Ana Sabina Uban .....	34
<i>Towards Preparation of the Second BUCC Shared Task: Detecting Parallel Sentences in Comparable Corpora</i> Pierre Zweigenbaum, Serge Sharoff and Reinhard Rapp .....	38





## Introduction to BUCC 2016

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 chiefly 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 emerging field by bundling its research, thereby making it more visible and giving it a better platform.

Following the eight previous editions of the workshop which took place in Africa (LREC’08 in Marrakech), America (ACL’11 in Portland), Asia (ACL-IJCNLP’09 in Singapore and ACL-IJCNLP’15 in Beijing), Europe (LREC’10 in Malta, ACL’13 in Sofia, and LREC’14 in Reykjavik) and also on the border between Asia and Europe (LREC’12 in Istanbul), the workshop this year is co-located with LREC’16 in Portorož, Slovenia.

We would like to thank all people who in one way or another helped in making this workshop once again a success. Our special thanks go to Ruslan Mitkov and Gregory Grefenstette for accepting to give invited presentations, to the members of the program committee who did an excellent job in reviewing the submitted papers under strict time constraints, and to the LREC’16 workshop chairs and organizers. Last but not least we would like to thank our authors and the participants of the workshop.

Reinhard Rapp, Pierre Zweigenbaum, Serge Sharoff

May 2016