# BUCC, 7th Workshop on Building and Using Comparable Corpora
# Building Resources for Machine Translation Research

Co-located with LREC 2014
Reykjavik (Iceland)
27 May 2014

http://comparable.limsi.fr/bucc2014/
Submission site:
https://www.softconf.com/lrec2014/BUCC2014/
Journal special issue: JNLE
Invited speaker: **Chris Callison-Burch**
University of Pennsylvania

## INVITED SPEAKER

**Chris Callison-Burch** University of Pennsylvania

## MOTIVATION

In the language engineering and the 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 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.

The scarcity of parallel corpora has motivated research concerning the use of comparable corpora: pairs of monolingual corpora selected according to the same set of criteria, but in different languages or language varieties. Non-parallel yet comparable corpora overcome the two limitations of parallel corpora, since sources for original, monolingual texts are much more abundant than translated texts. However, because of their nature, mining translations in comparable corpora is much more challenging than in parallel corpora. What constitutes a good comparable corpus, for a given task or per se, also requires specific attention: while the definition of a parallel corpus is fairly straightforward, building a non-parallel corpus requires control over the selection of source texts in both languages.

Parallel corpora are a key resource as training data for statistical machine translation, and for building or extending bilingual lexicons and terminologies. However, beyond a few language pairs such as English- French or English-Chinese and a few contexts such as parliamentary debates or legal texts, they remain a scarce resource, despite the creation of automated methods to collect parallel corpora from the Web. To exemplify such issues in a practical setting, this year's special focus will be on

Building Resources for Machine Translation Research

This special topic aims to address the need for:

1. Machine Translation training and testing data such as spoken or written monolingual, comparable or parallel data collections, and
2. Methods and tools used for collecting, annotating, and verifying MT data such as Web crawling, crowdsourcing, tools for language experts and for finding MT data in comparable corpora.

# TOPICS

We solicit contributions including but not limited to the following topics.

Topics related to the special theme:

- Methods and tools for collecting and processing MT data, including crowdsourcing
- Methods and tools for quality control
- Tools for efficient annotation
- Bilingual term and named entity collections
- Multilingual treebanks, wordnets, propbanks, etc.
- Comparable corpora with parallel units annotated
- Comparable corpora for under-resourced languages and specific domains
- Multilingual corpora with rich annotations: POS tags, NEs, dependencies, semantic roles, etc.
- Data for special applications: patent translation, movie subtitles, MOOCs, meetings, chatrooms, social media, etc.
- Legal issues with collecting and redistributing data and generating derivatives

Building Comparable Corpora:

- Human translations
- Automatic and semi-automatic methods
- 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, across language families
- Multi-media/multi-modal comparable corpora

Applications of comparable corpora:

- Human translations
- Language learning
- Cross-language information retrieval & document categorization
- Bilingual projections
- Machine translation

- Writing assistance

Mining from Comparable Corpora:

- Extraction of parallel segments or paraphrases from comparable corpora
- Extraction of bilingual and multilingual translations of single words and multi-word expressions; proper names, named entities, etc.

Note that an edited book "Building and Using Comparable Corpora" has just been published by Springer.

Chapter 1, an introduction and state of the art on the topic, is now freely available on Springer's Web site: Overviewing Important Aspects of the Last 20 Years of Research in Comparable Corpora.

## IMPORTANT DATES

| | |
|---|---|
| 23 February 2014 | Deadline for submission of full papers |
| 10 March 2014 | Notification of acceptance |
| 27 March 2014 | Camera-ready papers due |
| 27 May 2014 | Workshop date |

## SUBMISSION INFORMATION

Papers should follow the LREC main conference formatting details at http://lrec2014.lrec-conf.org/en/submission/autho
kit/ and should be submitted as a PDF-file via the START workshop manager at https://www.softconf.com/lrec2014/BUCC2014/.

Contributions can be short or long papers. Short paper submission must describe original and unpublished work without exceeding six (6) pages. Characteristics of short papers include: a small, focused contribution; work in progress; a negative result; an opinion piece; an interesting application nugget. Long paper submissions must describe substantial, original, completed and unpublished work without exceeding ten (10) pages.

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 immediately notified to the workshop organizers.

When submitting a paper from the START page, authors will be asked to provide essential information about resources (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 your research. Moreover, ELRA encourages all LREC authors to share the described LRs (data, tools, services, etc.), to enable their reuse, replicability of experiments, including evaluation ones, etc.

For further information, please contact Pierre Zweigenbaum mailto:pz(erase_at)limsi(erase_dot)fr

Plain-text CFP : bucc2014-cfp.txt
PDF CFP : bucc2014-cfp.pdf
Last modified: 12 Jul 2014

### JOURNAL SPECIAL ISSUE

Authors of selected papers will be encouraged to submit substantially extended versions of their manuscripts to an upcoming special issue on "Machine Translation Using Comparable Corpora" of the Journal of Natural Language Engineering.

## ORGANISERS

**Pierre Zweigenbaum** LIMSI, CNRS, Orsay (France), Chair

**Ahmet Aker** University of Sheffield (UK)

**Serge Sharoff** University of Leeds (UK)

**Stephan Vogel** QCRI (Qatar)

**Reinhard Rapp** Universities of Mainz (Germany) and Aix-Marseille (France)

## SCIENTIFIC COMMITTEE

Ahmet Aker, University of Sheffield (UK)
Srinivas Bangalore (AT&T Labs, US)
Caroline Barrière (CRIM, Montréal, Canada)
Chris Biemann (TU Darmstadt, Germany)
Hervé Déjean (Xerox Research Centre Europe, Grenoble, France)
Kurt Eberle (Lingenio, Heidelberg, Germany)
Andreas Eisele (European Commission, Luxembourg)
Éric Gaussier (Université Joseph Fourier, Grenoble, France)
Gregory Grefenstette (INRIA, Saclay, France)
Silvia Hansen-Schirra (University of Mainz, Germany)
Hitoshi Isahara (Toyohashi University of Technology)
Kyo Kageura (University of Tokyo, Japan)
Adam Kilgarriff (Lexical Computing Ltd, UK)
Natalie Kübler (Université Paris Diderot, France)
Philippe Langlais (Université de Montréal, Canada)
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 (Université Aix-Marseille, France)
Sujith Ravi (Google, US)
Serge Sharoff (University of Leeds, UK)
Michel Simard (National Research Council Canada)
Richard Sproat (OGI School of Science & Technology, US)
Tim Van de Cruys (IRIT-CNRS, Toulouse, France)
Stephan Vogel, QCRI (Qatar)
Guillaume Wisniewski (Université Paris Sud & LIMSI-CNRS, Orsay, France)
Pierre Zweigenbaum (LIMSI-CNRS, Orsay, France)