

# **Eighth Workshop on Building and Using Comparable Corpora 2015**

**Co-located with ACL-IJCNLP 2015**

**Beijing, China  
30 July 2015**

**ISBN: 978-1-5108-0946-8**

**Printed from e-media with permission by:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571



**Some format issues inherent in the e-media version may also appear in this print version.**

Copyright© (2015) by the Association for Computational Linguistics  
All rights reserved.

Printed by Curran Associates, Inc. (2015)

For permission requests, please contact the Association for Computational Linguistics  
at the address below.

Association for Computational Linguistics  
209 N. Eighth Street  
Stroudsburg, Pennsylvania 18360

Phone: 1-570-476-8006  
Fax: 1-570-476-0860

[acl@aclweb.org](mailto:acl@aclweb.org)

**Additional copies of this publication are available from:**

Curran Associates, Inc.  
57 Morehouse Lane  
Red Hook, NY 12571 USA  
Phone: 845-758-0400  
Fax: 845-758-2634  
Email: [curran@proceedings.com](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

## Table of Contents

<i>Augmented Comparative Corpora and Monitoring Corpus in Chinese: LIVAC and Sketch Search Engine Compared</i>	
Benjamin K. Tsou .....	1
<i>A Factory of Comparable Corpora from Wikipedia</i>	
Alberto Barrón-Cedeño, Cristina España-Bonet, Josu Boldoba and Lluís Màrquez .....	3
<i>Knowledge-lean projection of coreference chains across languages</i>	
Yulia Grishina and Manfred Stede .....	14
<i>Projective methods for mining missing translations in DBpedia</i>	
Laurent Jakubina and Philippe Langlais .....	23
<i>Attempting to Bypass Alignment from Comparable Corpora via Pivot Language</i>	
Alexis Linard, Béatrice Daille and Emmanuel Morin .....	32
<i>Application of a Corpus to Identify Gaps between English Learners and Native Speakers</i>	
Katsunori Kotani and Takehiko Yoshimi .....	38
<i>A Generative Model for Extracting Parallel Fragments from Comparable Documents</i>	
Somayeh Bakhshaei, Shahram Khadivi and Reza Safabakhsh .....	43
<i>Evaluating Features for Identifying Japanese-Chinese Bilingual Synonymous Technical Terms from Patent Families</i>	
Zi Long, Takehito Utsuro, Tomoharu Mitsuhashi and Mikio Yamamoto .....	52
<i>Extracting Bilingual Lexica from Comparable Corpora Using Self-Organizing Maps</i>	
Hyeong-Won Seo, Minah Cheon and Jae-Hoon Kim .....	62
<i>Obtaining SMT dictionaries for related languages</i>	
Miguel Rios and Serge Sharoff .....	68
<i>BUCC Shared Task: Cross-Language Document Similarity</i>	
Serge Sharoff, Pierre Zweigenbaum and Reinhard Rapp .....	74
<i>AUT Document Alignment Framework for BUCC Workshop Shared Task</i>	
Atefeh Zafarian, Amir Pouya Agha Sadeghi, Fatemeh Azadi, Sonia Ghiasifard, Zeinab Ali Panahloo, Somayeh Bakhshaei and Seyyed Mohammad Mohammadzadeh Ziabary .....	79
<i>LINA: Identifying Comparable Documents from Wikipedia</i>	
Emmanuel Morin, Amir Hazem, Florian Boudin and Elizaveta Loginova-Clouet .....	88