

# **Sixth Workshop on Building and Using Comparable Corpora 2013**

**Sofia, Bulgaria  
8 August 2013**

**ISBN: 978-1-62993-011-4**

**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© (2013) by the Association for Computational Linguistics  
All rights reserved.

Printed by Curran Associates, Inc. (2013)

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>Cross-lingual WSD for Translation Extraction from Comparable Corpora</i> Marianna Apidianaki, Nikola Ljubešić and Darja Fišer .....	1
<i>Bilingual Lexicon Extraction via Pivot Language and Word Alignment Tool</i> Hong-seok Kwon, Hyeong-won Seo and Jae-hoon Kim .....	11
<i>Using WordNet and Semantic Similarity for Bilingual Terminology Mining from Comparable Corpora</i> Dhouha Bouamor, Nasredine Semmar and Pierre Zweigenbaum .....	16
<i>A Comparison of Smoothing Techniques for Bilingual Lexicon Extraction from Comparable Corpora</i> Amir HAZEM and Emmanuel MORIN .....	24
<i>Chinese–Japanese Parallel Sentence Extraction from Quasi–Comparable Corpora</i> Chenhui Chu, Toshiaki Nakazawa and Sadao Kurohashi .....	34
<i>A modular open-source focused crawler for mining monolingual and bilingual corpora from the web</i> Vassilis Papavassiliou, Prokopis Prokopidis and Gregor Thurmair .....	43
<i>Building basic vocabulary across 40 languages</i> Judit Acs, Katalin Pajkossy and Andras Kornai .....	52
<i>Scientific registers and disciplinary diversification: a comparable corpus approach</i> Elke Teich, Stefania Degaetano-Ortlieb, Hannah Kermes and Ekaterina Lapshinova-Koltunski ..	59
<i>Improving MT System Using Extracted Parallel Fragments of Text from Comparable Corpora</i> Rajdeep Gupta, Santanu Pal and Sivaji Bandyopadhyay .....	69
<i>VARTRA: A Comparable Corpus for Analysis of Translation Variation</i> Ekaterina Lapshinova-Koltunski .....	77
<i>Building Ontologies from Collaborative Knowledge Bases to Search and Interpret Multilingual Corpora</i> Yegin Genc, Elizabeth Lennon, Winter Mason and Jeffrey Nickerson .....	87
<i>Using a Random Forest Classifier to recognise translations of biomedical terms across languages</i> Georgios Kontonatsios, Ioannis Korkontzelos, Sophia Ananiadou and Jun’ichi Tsujii .....	95
<i>Comparing Multilingual Comparable Articles Based On Opinions</i> Motaz Saad, David Langlois and Kamel Smaili .....	105
<i>Mining for Domain-specific Parallel Text from Wikipedia</i> Magdalena Plamada and Martin Volk .....	112
<i>Gathering and Generating Paraphrases from Twitter with Application to Normalization</i> Wei Xu, Alan Ritter and Ralph Grishman .....	121
<i>Learning Comparable Corpora from Latent Semantic Analysis Simplified Document Space</i> Ekaterina Stambolieva .....	129
<i>Finding More Bilingual Webpages with High Credibility via Link Analysis</i> Chengzhi Zhang, Xuchen Yao and Chunyu Kit .....	138