

12th Web as Corpus Workshop (ACL SIGWAC 2020)

Marseille, France
11 – 16 May 2020

Editors:

**Adrien Barbaresi
Felix Bildhauer**

**Roland Schafer
Egon Stemle**

ISBN: 978-1-7138-1251-7

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© (2020) by the Association for Computational Linguistics
All rights reserved.

Copyright for individual papers remains with the authors and are licensed under a Creative Commons 4.0 license, CC-BY-ND. (<https://creativecommons.org/licenses/by-nd/4.0/>)

Printed with permission by Curran Associates, Inc. (2020)

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

Additional copies of this publication are available from:

Curran Associates, Inc.
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: 845-758-0400
Fax: 845-758-2633
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

<i>Current Challenges in Web Corpus Building</i>	
Miloš Jakubíček, Vojtěch Kovář, Pavel Rychlý and Vit Suchomel	1
<i>Out-of-the-Box and into the Ditch? Multilingual Evaluation of Generic Text Extraction Tools</i>	
Adrien Barbaresi and Gaël Lejeune	5
<i>From Web Crawl to Clean Register-Annotated Corpora</i>	
Veronika Laippala, Samuel Rönqvist, Saara Hellström, Juhani Luotolahti, Liina Repo, Anna Salmela, Valtteri Skantsi and Sampo Pyysalo	14
<i>Building Web Corpora for Minority Languages</i>	
Heidi Jauhiainen, Tommi Jauhiainen and Krister Lindén	23
<i>The ELTE.DH Pilot Corpus – Creating a Handcrafted Gigaword Web Corpus with Metadata</i>	
Balázs Indig, Árpád Knap, Zsófia Sárközi-Lindner, Mária Timári and Gábor Palkó	33
<i>Hypernym-LIBre: A Free Web-based Corpus for Hypernym Detection</i>	
Shaurya Rawat, Mariano Rico and Oscar Corcho	42
<i>A Cross-Genre Ensemble Approach to Robust Reddit Part of Speech Tagging</i>	
Shabnam Behzad and Amir Zeldes	50
<i>Streaming Language-Specific Twitter Data with Optimal Keywords</i>	
Tim Kreutz and Walter Daelemans	57