9th Web as Corpus Workshop

(WaC-9 2014)

Held at the 14th Conference of the European Chapter
of the Association for Computational Linguistics

Gothenburg, Sweden
26 April 2014

ISBN: 978-1-63266-399-3

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© (2014) by the Association for Computational Linguistics
All rights reserved.

Printed by Curran Associates, Inc. (2014)

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: 845-758-2634
Email: curran@proceedings.com
Web: www.proceedings.com

Table of Contents

Finding Viable Seed URLs for Web Corpora: A Scouting Approach and Comparative Study of Available Sources
Adrien Barbaresi 1

Focused Web Corpus Crawling
Roland Schäfer, Adrien Barbaresi and Felix Bildhauer 9

Less Destructive Cleaning of Web Documents by Using Standoff Annotation
Maik Stührenberg 16

Some Issues on the Normalization of a Corpus of Products Reviews in Portuguese
Magali Sanches Duran, Lucas Avanço, Sandra Aluísio, Thiago Pardo and Maria da Graça Volpe Nunes 22

bs,hr,srWaC - Web Corpora of Bosnian, Croatian and Serbian
Nikola Ljubešić and Filip Klubička 29

The PAISÀ Corpus of Italian Web Texts
Verena Lyding, Egon Stemle, Claudia Borghetti, Marco Brunello, Sara Castagnoli, Felice Dell'Orletta, Henrik Dittmann, Alessandro Lenci and Vito Pirrelli 36

Internet Data in a Study of Language Change and a Program Helping to Work with Them
Varvara Magomedova, Natalia Slioussar and Maria Kholodilova . . . . .N/A