LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

The 12th Web as Corpus Workshop
(ACL SIGWAC)

PROCEEDINGS

Editors:
Adrien Barbaresi, Felix Bildhauer, Roland Schäfer and Egon Stemle

Proceedings of the LREC 2020
12th Web as Corpus Workshop
(ACL SIGWAC)

Edited by: Adrien Barbaresi, Felix Bildhauer, Roland Schäfer and Egon Stemle

ISBN: 979-10-95546-68-9
EAN: 9791095546689

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

Organizers:

Adrien Barbaresi, BBAW Berlin, DE
Felix Bildhauer, IDS Mannheim & Humboldt-Universität zu Berlin, DE
Roland Schäfer, Humboldt-Universität zu Berlin, DE
Egon Stemle, Eurac Research, IT

Program Committee:

Piotr Banski, IDS Mannheim, DE
Silvia Bernardini, University of Bologna, IT
Stefan Evert, University of Erlangen, DE
Miloš Jakubíček, SketchEngine, UK
Simon Krek, Jožef Stefan Institute, SI
Nikola Ljubešić, Jožef Stefan Institute, SI
Elizabeth Pankratz, University of Potsdam, DE
Steffen Remus, University of Hamburg, DE
Serge Sharoff, University of Leeds, UK
Wajdi Zaghouani, Hamad Bin Khalifa University, QA

Felix Bildhauer's and Roland Schäfer's work was partially funded by the Deutsche Forschungsgegemeinschaft (German Research Foundadtion) – SFB 1412 Register – 416591334

Introduction

For almost fifteen years, the ACL SIGWAC, and most notably the Web as Corpus (WAC) workshops, have served as a platform for researchers interested in the compilation, processing and use of web-derived corpora as well as computer-mediated communication. Past workshops were co-located with major conferences on corpus linguistics and computational linguistics (such as ACL, EACL, Corpus Linguistics, LREC, NAACL, WWW).

In corpus linguistics and theoretical linguistics, the World Wide Web has become increasingly popular as a source of linguistic evidence, especially in the face of data sparseness or the lack of variation in traditional corpora of written language. In lexicography, web data have become a major and well-established resource with dedicated research data and specialised tools. In other areas of theoretical linguistics, the adoption rate of web corpora has been slower but steady. Furthermore, some completely new areas of linguistic research dealing exclusively with web (or similar) data have emerged, such as the construction and utilisation of corpora based on short messages. Another example is the (manual or automatic) classification of web texts by genre, register, or – more generally speaking – "text type", as well as topic area. In computational linguistics, web corpora have become an established source of data for the creation of language models, word embeddings, and all types of machine learning.

The 12th Web as Corpus workshop (WAC-XII) looks at the past, present, and future of web corpora given the fact that large web corpora are nowadays provided mostly by a few major initiatives and companies, and the diversity of the early years appears to have faded slightly. Also, we acknowledge the fact that alternative sources of data (such as data from Twitter and similar platforms) have emerged, some of them only available to large companies and their affiliates, such as linguistic data from social media and other forms of the deep web. At the same time, gathering interesting and relevant web data (web crawling) is becoming an ever more intricate task as the nature of the data offered on the web changes (for example the death of forums in favour of more closed platforms).

This year's edition will not lead to a half-day workshop at the LREC 2020 conference as expected. The full proceedings remain, of which the present volume is a part.

We received 13 submissions in total, the proceedings comprise 8 articles.

We would like to thank the reviewers for their help and wish to see you at the next edition under better circumstances.

The organizers:

Adrien Barbaresi
Felix Bildhauer
Roland Schäfer
Egon Stemle

Table of Contents