The 31st International Conference on Computational Linguistics

WACL-4

The 4th Workshop on Arabic Corpus Linguistics

Proceedings of the Workshop

January 20, 2025

https://wp.lancs.ac.uk/wacl4

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
209 N. Eighth Street
Stroudsburg, PA 18360
USA
Tel: +1-570-476-8006
Fax: +1-570-476-0860
acl@aclweb.org

ISBN 979-8-89176-220-6

Preface

Welcome to the Fourth Workshop on Arabic Corpus Linguistics (WACL-4), held online on January 20, 2025, in conjunction with the 31st International Conference on Computational Linguistics (COLING 2025) in Abu Dhabi, UAE.

The field of Arabic language research using corpora and corpus-based methods has undergone remarkable growth over the past decade. What began as a series of isolated initiatives has evolved into a dynamic and rapidly expanding domain of inquiry, encompassing a wide range of topics in both corpus and computational linguistics. Building on the success of the previous workshops—WACL-1 (2011), WACL-2 (2013, hosted at the Corpus Linguistics Conference at Lancaster University), and WACL-3 (2019, hosted at the Corpus Linguistics Conference at Cardiff University)—WACL-4 (2025, hosted at COLING) continues to provide a dedicated venue for advancing research and promoting collaboration in this vibrant field.

The primary objectives of WACL-4 are to showcase the latest developments in the creation, annotation, and application of Arabic corpora and to foster interdisciplinary collaboration. This year, we place a special emphasis on Arabic dialects, including non-standard and regional varieties, aiming to deepen our understanding of Arabic in its many forms and to support research on under-resourced linguistic varieties. The workshop also seeks to encourage advancements in Natural Language Processing (NLP) tailored for Arabic, focusing on integrating corpora into NLP workflows, developing new computational tools, and evaluating existing systems to enhance their performance in processing Arabic text.

We received 22 submissions most of which 13 were accepted. Each submission underwent rigorous review by at least three reviewers, ensuring the quality and relevance of the accepted contributions, resulting in an acceptance rate of 59

We thank the authors, reviewers, and organizing committee for their efforts and support. We hope these proceedings inspire new research and collaborations to advance the field.

Saad Ezzini, General Chair, on behalf of the organizing committee of the WACL-4 workshop.

Organizing Committee

Saad Ezzini, King Fahd University of Petroleum and Minerals, Saudi Arabia (General Chair)
Hamza Alami, Sidi Mohamed Ben Abdellah University, Morocco (Programme Co-Chair)
Ismail Berrada, Mohammed VI Polytechnic University, Morocco (Programme Co-Chair)
Abdessamad Benlahbib, Sidi Mohamed Ben Abdellah University, Morocco (Programme Co-Chair)
Abdelkader El Mahdaouy, Mohammed VI Polytechnic University, Morocco (Review Chair)
Salima Lamsiyah, University of Luxembourg, Luxembourg (Publication Chair)
Hatim Derrouz, Ibn Tofail University, Morocco (Publicity Co-Chair)
Amal Haddad, University of Granada, Spain (Publicity Co-Chair)
Mustafa Jarrar, Birzeit University, Palestine (Advisory Committee)
Mo El-Haj, Lancaster University, UK (Advisory Committee)
Ruslan Mitkov, Lancaster University, UK (Advisory Committee)
Paul Rayson, Lancaster University, UK (Advisory Committee)

Programme Committee

Almoataz B. Al-Said, Cairo University, Egypt
Abdessamad Benlahbib, Sidi Mohamed Ben Abdellah University, Morocco
Ashraf Boumhidi, Sidi Mohamed Ben Abdellah University, University, Morocco
Abdelkader El Mahdaoui, Mohammed VI Polytechnic University, Morocco
Hamza Alami, Sidi Mohamed Ben Abdellah University, Morocco
Hatim Derrouz, Ibn Tofail University, Morocco
Hicham Hammouchi, University of Luxembourg, Luxembourg
Ismail Berrada, Mohammed VI Polytechnic University, Morocco
Maram Alharbi, Lancaster University, UK
Nagham F. Hamad, Birzeit University, Palestine
Nizar Habash, New York University Abu Dhabi, UAE
Nora Al-Twairesh, King Saud University, Saudi Arabia
Noorhan Abbas, Leeds University, UK
Saad Ezzini, King Fahd University of Petroleum and Minerals, Saudi Arabia
Salima Lamsiyah, University of Luxembourg, Luxembourg
Salmane Chafik, Mohammed VI Polytechnic University, Morocco
Samir El Amrani, University of Luxembourg, Luxembourg
Wajdi Zaghouani, Hamad Bin Khalifa University, Qatar

Table of Contents

Conference Program

Monday, January 20, 2025

9:00–9:10 ***Welcome and Opening Remarks***

9:10–9:50 *Invited Talk by Imed Zitouni: Bridging the Gap: Arabic Search in the Age of LLMs*

Session 1

9:50–10:10 *ArabicSense: A Benchmark for Evaluating Commonsense Reasoning in Arabic with Large Language Models*
Salima Lamsiyah, Kamyar Zeinalipour, Samir El Amrany, Matthias Brust, Marco Maggini, Pascal Bouvry and Christoph Schommer

10:10–10:30 *Lahjawi: Arabic Cross-Dialect Translator*
Mohamed Motasim Hamed, Muhammad Hreden, Khalil Hennara, Zeina Aldallal, Sara Chrouf and Safwan AlModhayan

10:30–11:00 ***Coffee Break***

Session 2

11:00–11:20 *Lost in Variation: An Unsupervised Methodology for Mining Lexico-syntactic Patterns in Middle Arabic Texts*
Julien JB Bezançon, Rimane Karam and Gaël Lejeune

11:20–11:40 *SADSLyC: A Corpus for Saudi Arabian Multi-dialect Identification through Song Lyrics*
Salwa Saad Alahmari

11:40–12:00 *Enhancing Dialectal Arabic Intent Detection through Cross-Dialect Multilingual Input Augmentation*
Shehenaz Hossain, Fouad Shammary, Bahaulddin Shammary and Haithem Afli

12:00–12:20 *Dial2MSA-Verified: A Multi-Dialect Arabic Social Media Dataset for Neural Machine Translation to Modern Standard Arabic*
Abdullah Salem Khered, Youcef Benkhedda and Riza Batista-Navarro

12:20–13:20 ***Lunch Break***

Monday, January 20, 2025 (continued)

Session 3

13:20–13:40 *Web-Based Corpus Compilation of the Emirati Arabic Dialect*
Yousra A. El-Ghawi

13:40–14:00 *Evaluating Calibration of Arabic Pre-trained Language Models on Dialectal Text*
Ali Al-Laith and Rachida Kebdani

14:00–14:20 *Empirical Evaluation of Pre-trained Language Models for Summarizing Moroccan Darija News Articles*
Azzedine Aftiss, Salima Lamsiyah, Christoph Schommer and Said Ouatik El Alaoui

14:20–14:40 *Dialect2SQL: A Novel Text-to-SQL Dataset for Arabic Dialects with a Focus on Moroccan Darija*
Salmane Chafik, Saad Ezzini and Ismail Berrada

14:40–15:00 *AraSim: Optimizing Arabic Dialect Translation in Children's Literature with LLMs and Similarity Scores*
Alaa Hassan Bouomar and Noorhan Abbas

15:00–15:20 *Navigating Dialectal Bias and Ethical Complexities in Levantine Arabic Hate Speech Detection*
Ahmed Haj Ahmed, Rui-Jie Yew, Xerxes Minocher and Suresh Venkatasubramanian

15:20–16:00 ***Coffee Break***

16:00–16:30 ***Best Paper Award, Closing Remarks, and Wrap-Up by Dr Saad Ezzini***