LREC-COLING 2024

Proceedings of the Second Workshop on Computation and Written Language (CAWL 2024) @LREC-COLING-2024

Workshop Proceedings

Editors
Kyle Gorman

May 21, 2024
Torino, Italia

Proceedings of Second Workshop on Computation and Written Language (CAWL 2024) @LREC-COLING-2024

ISBN 978-2-493814-28-9
ISSN 2951-2093 (COLING); 2522-2686 (LREC)

Jointly organized by the ELRA Language Resources Association
and the International Committee on Computational Linguistics

Message from the Chairs

Welcome to the second meeting of the CAWL workshop, featuring eight original paper presentations, an invited talk by Nizar Habash, and an invited lecture by Jalal Maleki.

This year's workshop is sponsored by Google; we thank them for their support.

Since our first meeting in Toronto in July 2023, we have formed a special interest group: the ACL Special Interest Group on Writing Systems and Written Language, or SIGWrit for short. This SIG will be responsible for organizing future meetings of CAWL, and may pursue other objectives as decided by the officers and members of the SIG.

The current *pro tempore* officers of SIGWrit are president Richard Sproat, vice president Emily Prud'hommeaux, secretary-treasurer Kyle Gorman, and student member Noah Hermalin. As per ACL policy, we will organize an election for new officers in fall 2024. More information about the SIG, including its constitution, can be found at `https://sigwrit.org/`.

Invited Talks and Lectures

Nizar Habash: On Writing Arabic

The Arabic language, broadly defined, encompasses a diverse collection of varieties that are tied together historically and linguistically, but with a high degree of variations in terms of phonology, morphology, lexicon, and naturally orthography. In this talk we present a condensed summary of the challenges of writing Arabic and the evolution of different orthographic solutions to address them. The accumulation and persistence of different conventions have led to many co-existing orthographies today creating a complex space of challenges for computational modeling. Among the examples we discuss are subtle differences in Standard Arabic spelling across Arab countries, using scripts other than Arabic for writing Arabic dialects, and, most recently, social media experimentation with reverting to ancient orthographic conventions to fight AI censorship algorithms.

Jalal Maleki: Balancing Linguistic Integrity and Practicality: The Design Journey of Dabire, a Romanized Writing System for Persian

Developing a new writing system, like the romanized Dabire for Persian, requires a nuanced balance between adhering to orthographic principles and making pragmatic compromises. At the core of Dabire's design is the principle of linguistic soundness, with phonemicity as a cornerstone, ensuring a direct, systematic encoding of sounds to simplify the learning process and enhance teaching efficacy. The focus on phonemicity, crucial for the script's ease of use and learning, particularly benefits young learners and non-native speakers. However, the design process also involves balancing phonemic and morphophonemic considerations, acknowledging the complex interplay between adjacent morphemes on sound realization. In practice, rigorous maintenance of both phonemicity and morphophonemic consistency is impossible and concessions are sometimes necessary. Other properties of the Dabire writing system that will be discussed are faithfulness, transparency, completeness and orthographic depth. This talk will highlight the considerations and compromises in designing Dabire, revealing the challenges and opportunities of developing a practical, efficient, and linguistically sound writing system for Persian.

Organizing Committee

- Kyle Gorman, Graduate Center, City University of New York and Google, USA
- Emily Prud'hommeaux, Boston College, USA
- Brian Roark, Google, USA
- Richard Sproat, Google DeepMind, Japan

Program Committee

- David Ifeoluwa Adelani
- Manex Agirrezabal
- Sina Ahmadi
- Cecilia Alm
- Mark Aronoff
- Steven Bedrick
- Taylor Berg-Kirkpatrick
- Amalia Gnanadesikan
- Christian Gold
- Alexander Gutkin
- Nizar Habash
- Yannis Haralambous
- Cassandra Jacobs
- Martin Jansche
- Kathryn Kelley
- George Kiraz
- Christo Kirov
- Jordan Kodner
- Anoop Kunchukuttan
- Yang Li
- Constantine Lignos
- Zoey Liu
- Jalal Maleki
- M. Willis Monroe
- Gerald Penn
- Yuval Pinter
- William Poser
- Shruti Rijhwani
- Maria Ryskina
- Anoop Sarkar
- Lane Schwartz
- Djamé Seddah
- Shuming Shi
- Claytone Sikasote
- Fabio Tamburini
- Kumiko Tanaka-Ishii
- Lawrence Wolf-Sonkin
- Martha Yifiru Tachbelie

Table of Contents

Tutorial Program

Tuesday, May 21, 2024

09:00–09:10	*Opening remarks* Organizers
09:10–10:10	*Invited talk: On Writing Arabic* Nizar Habash
10:10–10:30	*ParsText: A Digraphic Corpus for Tajik-Farsi Transliteration* Rayyan Merchant and Kevin Tang
10:30–11:00	***Coffee break***
11:30–12:00	*A Joint Approach for Automatic Analysis of Reading and Writing Errors* Wieke Harmsen, Catia Cucchiarini, Roeland van Hout and Helmer Strik
12:00–12:20	*Tool for Constructing a Large-Scale Corpus of Code Comments and Other Source Code Annotations* Luna Peck and Susan Brown
12:20–14:00	***Lunch break***
14:00–14:30	*Tokenization via Language Modeling: the Role of Preceding Text* Rastislav Hronsky and Emmanuel Keuleers
14:30–14:50	*Abbreviation Across the World's Languages and Scripts* Kyle Gorman and Brian Roark
14:50–15:20	*Now You See Me, Now You Don't: 'Poverty of the Stimulus' Problems and Arbitrary Correspondences in End-to-End Speech Models* Daan van Esch
15:20–15:40	*Towards Fast Cognate Alignment on Imbalanced Data* Logan Born, M. Willis Monroe, Kathryn Kelley and Anoop Sarkar
15:40–16:00	*Business meeting* Organizers

Tuesday, May 21, 2024 (continued)

16:00–16:30 ***Coffee break***

16:30–16:50 *Simplified Chinese Character Distance Based on Ideographic Description Sequences*
Yixia Wang and Emmanuel Keuleers