Workshop on Computation and Written Language (CAWL 2023)

Toronto, Canada
14 July 2023

ISBN: 978-1-7138-8223-7

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© (2023) by the Association for Computational Linguistics
All rights reserved.

Printed with permission by Curran Associates, Inc. (2024)

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

Myths about Writing Systems in Speech & Language Technology
Kyle Gorman and Richard Sproat 1

The Hidden Folk: Linguistic Properties Encoded in Multilingual Contextual Character Representations
Manex Agirrezabal, Sidsel Boldsen and Nora Hollenstein 6

Preserving the Authenticity of Handwritten Learner Language: Annotation Guidelines for Creating Transcripts Retaining Orthographic Features
Christian Gold, Ronja Laarmann-quante and Torsten Zesch 14

Exploring the Impact of Transliteration on NLP Performance for Low-Resource Languages: The Case of Maltese and Arabic
Kurt Micallef, Fadhl Eryani, Nizar Habash, Houda Bouamor and Claudia Borg 22

Distinguishing Romanized Hindi from Romanized Urdu
Elizabeth Nielsen, Christo Kirov and Brian Roark 33

Back-Transliteration of English Loanwords in Japanese
Yuying Ren 43

Pronunciation Ambiguities in Japanese Kanji
Wen Zhang 50

Lenient Evaluation of Japanese Speech Recognition: Modeling Naturally Occurring Spelling Inconsistency
Shigeki Karita, Richard Sproat and Haruko Ishikawa 61

Disambiguating Numeral Sequences to Decipher Ancient Accounting Corpora
Logan Born, M. Willis Monroe, Kathryn Kelley and Anoop Sarkar 71

Decipherment of Lost Ancient Scripts as Combinatorial Optimisation Using Coupled Simulated Annealing
Fabio Tamburini 82

Learning the Character Inventories of Undeciphered Scripts Using Unsupervised Deep Clustering
Logan Born, M. Willis Monroe, Kathryn Kelley and Anoop Sarkar 92

A Mutual Information-based Approach to Quantifying Logography in Japanese and Sumerian
Noah Hermalin 105