CAWL 2023

The Workshop on Computation and Written Language (CAWL)

Proceedings of the Workshop

July 14, 2023

The CAWL organizers gratefully acknowledge the support from the following sponsors.

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 978-1-959429-90-6

Introduction

We are pleased to bring you these Proceedings of the Workshop on Computation and Written Language (CAWL), held in Toronto on July 14, 2023. We received 17 paper submissions, of which 11 were chosen to appear in the workshop. Additionally, we include a position paper from organizers and the abstracts of our two invited talks.

Most work on NLP focuses on language in its canonical written form. This has often led researchers to ignore the differences between written and spoken language or, worse, to conflate the two. Instances of conflation are statements like "Chinese is a logographic language" or "Persian is a right-to-left language", variants of which can be found frequently in the ACL anthology. These statements confuse properties of the language with properties of its writing system. Ignoring differences between written and spoken language leads, among other things, to conflating different words that are spelled the same, or treating as different, words that have multiple spellings.

Furthermore, methods for dealing with written language issues (e.g., various kinds of normalization or conversion) or for recognizing text input (e.g., OCR & handwriting recognition or text entry methods) are often regarded as precursors to NLP rather than as fundamental parts of the enterprise, despite the fact that most NLP methods rely centrally on representations derived from text rather than (spoken) language. This general lack of consideration of writing has led to much of the research on such topics to largely appear outside of ACL venues, in conferences or journals of neighboring fields such as speech technology (e.g., text normalization) or human-computer interaction (e.g., text entry).

We are excited to bring together researchers working on various aspects of these topics, and hope that this might be the means for creating a persistant community within ACL focused on these topics.

We would like to thank the members of the Program Committee for completing their reviews promptly, and for providing useful feedback for deciding on the program and preparing the final versions of the papers. Special thanks to Nizar Habash for helping with organizational issues. Thanks also to our invited speakers, Mark Aronoff and Amalia Gnanadesikan, and to the authors of the interesting papers we are presenting in this volume.

Kyle Gorman, Brian Roark, and Richard Sproat
Organizers of the workshop

Program Committee

Organizers

Kyle Gorman, The Graduate Center, City University of New York
Brian Roark, Google Inc.
Richard Sproat, Google, Japan

Program Committee

Manex Agirrezabal, University of Copenhagen
Sina Ahmadi, George Mason University
Cecilia Ovesdotter Alm, Rochester Institute of Technology
Steven Bedrick, Oregon Health & Science University
Taylor Berg-kirkpatrick, University of California San Diego
Dan Garrette, Google Research
Alexander Gutkin, Google
Nizar Habash, New York University Abu Dhabi
Yannis Haralambous, IMT Atlantique & CNRS LabSTICC
Cassandra L. Jacobs, University at Buffalo
Martin Jansche, Amazon
George Kiraz, Institute for Advanced Study
Christo Kirov, Google
Grzegorz Kondrak, University of Alberta
Yang Li, Northwestern Polytechnical University
Constantine Lignos, Brandeis University
Zoey Liu, Department of Linguistics, University of Florida
Gerald Penn, University of Toronto
Yuval Pinter, Ben-Gurion University of the Negev
William Poser, freelance
Emily Prud'hommeaux, Boston College
Shruti Rijhwani, Google
Maria Ryskina, Massachusetts Institute of Technology
Lane Schwartz, University of Alaska Fairbanks
Djamé Seddah, Inria
Shuming Shi, Tencent AI Lab
David Smith, Northeastern University
Kumiko Tanaka-ishii, Waseda University
Annalu Waller, University of Dundee

Invited Talks

Paradise Lost: How the Alphabet Fell from Perfection

Mark Aronoff
Stony Brook University

Abstract: The original alphabet, devised by Semitic speakers in Egypt ca. 1800 BCE, was a perfect 1-1 mapping between individual letters and individual sounds. All alphabets and similar systems are descended from this original invention. Very few alphabets today retain a perfect 1-1 mapping. How far have alphabets diverged from perfection since? Modern alphabetic systems have letter-to-phoneme mappings of up to 4-1. These included Italian (one of the most regular) and English (perhaps the least regular). English also shows a high number of mappings to single morphs (stems and affixes). Korean, the only alphabetic system that groups letters into syllables, shows an interaction between morphemes and syllabic grouping.

How Linguistic are Writing Systems?

Amalia Gnanadesikan
University of Maryland

Abstract: Theoretical linguists have long denied that writing is language, while NLP research has tended to conflate writing and spoken language. The truth is more complex than either view. Writing imposes a linguistic analysis on spoken language, dividing a continuous speech stream into segments, syllables, morphemes and/or words. These elements are not simply representational. Writing systems impose their own linguistic structure, for example by requiring syllables to meet well-formedness conditions even when the spoken syllables violate these conditions. Writing systems also include linguistic categories that are not used in the languages for which they are designed, such as graphic classifiers used in writing non-classifier languages or graphic inflectional morphology used for languages with virtually no inflectional morphology.

Table of Contents

Program

Friday, July 14, 2023

09:00 - 09:05 *Opening Remarks,* Organizers

09:05 - 09:15 *Position Paper*

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

09:15 - 10:15 *Invited talk, Mark Aronoff: Paradise Lost: How the Alphabet Fell from Perfection*

10:15 - 10:30 *Morning Talks*

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

10:30 - 11:00 *Coffee Break*

11:00 - 12:00 *Morning talks (continued)*

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

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

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

12:00 - 13:30 *Lunch Break*

13:30 - 14:30 *Invited Talk, Amalia Gnanadesikan: How Linguistic are Writing Systems?*

14:30 - 15:25 *Afternoon talks*

Friday, July 14, 2023 (continued)

Back-Transliteration of English Loanwords in Japanese
Yuying Ren

Pronunciation Ambiguities in Japanese Kanji
Wen Zhang

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

15:25 - 16:00 *Coffee Break*

16:00 - 17:15 *Afternoon talks (continued)*

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

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

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

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

17:15 - 17:30 *Closing Remarks,* Organizers