



LREC 2026

**The Third Workshop on
Computation and Written Language (CAWL 2026)
@LREC 2026**

Workshop Proceedings

**Editor
Kyle Gorman**

12 May, 2026

Preface

Welcome to the third meeting of the CAWL workshop, featuring an invited talk by Prof. Zev Handel of the University of Washington.

This is the second meeting of CAWL under the aegis of ACL Special Interest Group on Writing Systems and Written Language (SIGWrit), which will continue to organize future CAWL meetings. The present officers of SIGWrit are president Kyle Gorman, vice president Constantine Lignos, and secretary-treasurer Zoey Liu, and student member Claytone Sikasote. For more information about the SIG, see <https://sigwrit.org/>.

Organizing Committee

Kyle Gorman, CUNY Graduate Center and Google Research
Constantine Lignos, Brandeis University
Zoey Liu, University of Florida
Claytone Sikasote, University of Cape Town

Program Committee

David Ifeoluwa Adelani, University College London
Sina Ahmadi, George Mason University
Enas Albasiri, Nvidia Corporation
Cecilia Overdotter Alm, Rochester Institute of Technology
Steven Bedrick, Oregon Health & Science University
Alexander Gutkin, Google Research
Nizar Habash, NYU Abu Dhabi
Yannis Haralambous, IMT Atlantique and CNRS Lab-STICC
Christo Kirov, Google Research
Yuval Pinter, Ben-Gurion University of the Negev
Djamé Seddah, Sorbonne University and Inria
Brian Roark, Google Research
Maria Ryskina, Massachusetts Institute of Technology

Table of Contents

<i>The Degree of Language Diacriticity and Its Effect on Tasks</i> Adi Cohen and Yuval Pinter	1
<i>Private-Use Area Characters in the Wild: Signal or Noise?</i> Alexander Gutkin, Adrian Benton, Christo Kirov, Brian Roark and Lawrence Wolf-Sonkin	9
<i>HAnnol: A Handwriting Annotation Interface to Extract Data for Linguistic Analyses of Graphetic Detail</i> Joshua Wieler, Simon Petitjean, Kristian Berg, Henriette Huber and Stefan Hartmann	33
<i>SoriGraph: A New Database of Visual Feature-Level Descriptions of Written Korean</i> Wednesday Bushong, Hala Hababbeh, Ryan Jiang and Yoolim Kim	45
<i>Confusable Characters as Endangered Language Markers: The Case of North Caucasus Writing Systems</i> Alexander Gutkin, Adrian Benton, Christo Kirov and Brian Roark	50
<i>Prompting Approaches to Abbreviation Expansion</i> Kyle Gorman	64
<i>Evaluating Data Augmentation Strategies for Training Spanish Misspelling Detection Models</i> Manuel Castillo-Sancho, Jordi Porta and Asunción Gómez-Pérez	71
<i>Large Language Model-Based Post-OCR Correction for Low-Resource Kazakh Scripts</i> Henry Gagnier	79
<i>G&P2P: A Multi-Source Approach to Grapheme-to-Phoneme Conversion</i> Chun-Yi Jerry Peng	89
<i>A Lightweight N-gram Approach to Abbreviation Expansion in Large Corpora</i> Tjaša Šoltes and Marko Bajec	95
<i>Inverse Text Normalization for Arabic Numbers in Streaming ASR</i> Enas Albasiri, Myungjong Kim, Nourchene Ferchichi and Oluwatobi Olabiyi	101

Workshop Program

9:00–9:30	<i>Opening remarks</i> Kyle Gorman
9:30–10:30	<i>Everything you wanted to know about East Asian writing but didn't think to ask: The history and structure of scripts for Chinese, Japanese, Korean, and Vietnamese</i> Zev Handel
11:00–11:15	<i>The Degree of Language Diacriticity and Its Effect on Tasks</i> Adi Cohen and Yuval Pinter
11:15–11:30	<i>Private-Use Area Characters in the Wild: Signal or Noise?</i> Alexander Gutkin, Adrian Benton, Christo Kirov, Brian Roark and Lawrence Wolf-Sonkin
11:30–11:45	<i>HAnnot: A Handwriting Annotation Interface to Extract Data for Linguistic Analyses of Graphetic Detail</i> Joshua Wieler, Simon Petitjean, Kristian Berg, Henriette Huber and Stefan Hartmann
11:45–12:00	<i>SoriGraph: A New Database of Visual Feature-Level Descriptions of Written Korean</i> Wednesday Bushong, Hala Hababbeh, Ryan Jiang and Yoolim Kim
12:00–13:00	<i>Confusable Characters as Endangered Language Markers: The Case of North Caucasus Writing Systems</i> Alexander Gutkin, Adrian Benton, Christo Kirov and Brian Roark
12:00–13:00	<i>Prompting Approaches to Abbreviation Expansion</i> Kyle Gorman
12:00–13:00	<i>Evaluating Data Augmentation Strategies for Training Spanish Misspelling Detection Models</i> Manuel Castillo-Sancho, Jordi Porta and Asunción Gómez-Pérez
12:00–13:00	<i>Large Language Model-Based Post-OCR Correction for Low-Resource Kazakh Scripts</i> Henry Gagnier
12:00–13:00	<i>G&P2P: A Multi-Source Approach to Grapheme-to-Phoneme Conversion</i> Chun-Yi Jerry Peng
12:00–13:00	<i>A Lightweight N-gram Approach to Abbreviation Expansion in Large Corpora</i> Tjaša Šoltes and Marko Bajec

12:00–13:00

Inverse Text Normalization for Arabic Numbers in Streaming ASR

Enas Albasiri, Myungjong Kim, Nourchene Ferchichi and Oluwatobi Olabiyi