ACL 2023

The 20th SIGMORPHON workshop on Computational Morphology, Phonology, and Phonetics

July 14, 2023

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 978-1-959429-93-7

Introduction

eWelcome to the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology, to be held on July 14, 2023 as part of ACL in Toronto. The workshop aims to bring together researchers interested in applying computational techniques to problems in morphology, phonology, and phonetics. Our program this year highlights the ongoing investigations into how neural models process phonology and morphology, as well as the development of finite-state models for low-resource languages with complex morphology.

We received 22 submissions, and after a competitive reviewing process, we accepted 12, for an acceptance rate of 54.5%. The workshop is very happy to present two invited talks this year. Carmen Saldana, from the University of Zürich, and CUNY's Kyle Gorman presented talks at this year's workshop.

This year also marks the seventh iteration of the SIGMORPHON Shared Task. We hosted two Shared Tasks this year:
The UniMorph Shared Task on Typologically Diverse and Acquisition-Inspired Morphological Inflection Generation continued SIGMORPHON's tradition of shared tasks that investigate inflectional patterns. The task had two parts. The first part invited participants to build models that predict an inflected form from either a lemma, or other inflected form, as well as desired properties of the output. The second part investigates the cognitive plausibility of inflectional forms - namely, the task asks users to train a classification model that determines the phonological constraints that lead to generalization patterns in Korean; the final part investigates child-like errors made by inflectional systems.

The Shared Task on Interlinear glossing challenges participants to automate the process of glossing morphological processes in lower-resource languages - a task that is essential in language documentation. In the open track, participants train a model that produces a morphologically-specified gloss from the original source sentence, a canonically-segmented representation, and optionally, a second language translation. In the closed track, the segmented representation is absent.

We also present the results from the 2022 Shared Task on Cross-Lingual and Low-Resource Grapheme-Phoneme prediction. Due to time constraints with last year's proceedings, we were unable to publish the results. We apologize to the organizers and participants, who have had to wait a year to see their work in print.

We are grateful to the program committee for their careful and thoughtful reviews of the papers submitted this year. Likewise, we are thankful to the shared task organizers for their hard work in preparing the shared tasks. We are looking forward to a workshop covering a wide range of topics, and we hope for lively discussions.

Garrett Nicolai, Eleanor Chodroff, Çagri Çöltekin, and Fred Mailhot, workshop organization team.

Organizing Committee

Co-Chair

Garrett Nicolai, University of British Columbia
Eleanor Chodroff, University of York
Çagri Çöltekin, University of Tübingen
Fred Mailhot, Dialpad, Inc.

SIGMORPHON Officers

President: Garrett Nicolai, University of British Columbia
Secretary: Miikka Silfverberg, University of British Columbia
At Large: Eleanor Chodroff, University of York
At Large: Çağrı Çöltekin, University of Tübingen
At Large: Fred Mailhot, Dialpad, Inc.

Program Committee

Program Committee

Khuyagbaatar Batsuren, National University of Mongolia
Gasper Begus, University of California, Berkeley
Canaan Breiss, UCLA
Basilio Calderone, Université Toulouse Jean Jaurès & CNRS
Daniel Dakota, Indiana University
Aniello De Santo, University of Utah
Indranil Dutta, Jadavpur University
Jason Eisner, Johns Hopkins University + Microsoft Corporation
Micha Elsner, The Ohio State University
Michael Ginn, University of Colorado
Omer Goldman, Bar Ilan University
Nizar Habash, New York University Abu Dhabi
Nabil Hathout, CLLE, CNRS & Universite de Toulouse
Mathilde Hutin, Université Paris-Saclay, CNRS, LIMSI
Cassandra L. Jacobs, University at Buffalo
Adam Jardine, Rutgers University
Jordan Kodner, Stony Brook University
Sandra Kübler, Indiana University
Giorgio Magri, Centre National de la Recherche Scientifique
Rob Malouf, San Diego State University
Connor Mayer, UCLA
Sarah Moeller, University of Colorado
Kemal Oflazer, Carnegie Mellon University
Jeff Parker, Brigham Young University
Jelena Prokic, Leiden University
Jonathan Rawski, San Jose State University
Brian Roark, Google Inc.
Maria Ryskina, Massachusetts Institute of Technology
Miikka Silfverberg, University of British Columbia
Kairit Sirts, University of Tartu
Caitlin Smith, University of North Carolina at Chapel Hill
Morgan Sonderegger, McGill University
Kenneth Steimel, Educational Testing Service
Ekaterina Vylomova, University of Melbourne
Adam Wiemerslage, University of Colorado Boulder
Adina Williams, Meta Platforms, Inc.
Colin Wilson, Johns Hopkins University
Changbing Yang, University of British Columbia
Kristine Yu, University of Massachusetts Amherst

Keynote Talk: Cross-linguistic recurrent patterns in morphology mirror human cognition

Carmen Saldana
University of Zürich
2023-07-14 08:30:00 –

Abstract: A foundational goal of language science is to detect and define the set of constraints that explain cross-linguistic recurrent patterns (i.e., typological universals) in terms of fundamental shared features of human cognition. In this talk, I will present a series of Artificial Language Learning experimental studies which test a hypothesised link between biases in language learning and morphological universals in typology both at the syntagmatic (i.e., morpheme order) and paradigmatic levels (e.g., structure of inflectional paradigms). I will focus in particular on two types of universals in inflectional morphology: (1) affixes with stronger structural relationships to the word stem tend to appear linearly closer to it, and (2) different categories with the same identity (be it the same word form, or the same word structure) in morphological paradigms tend to be semantically similar. The results from the studies I will present provide evidence in favour of a shared typological and learning bias towards compositional transparency and locality in morpheme order, and a bias towards partitions of morphological paradigms that reflect semantic relatedness. In light of these results, I will argue that cross-linguistic recurrent morphological patterns mirror to some extend universal features of human cognition.

Bio: Carmen Saldana is currently a postdoctoral fellow in the Department of Comparative Language Science at the University of Zurich. Her research focuses on investigating the cognitive biases and processes that shape the current cross-linguistic distributions of morphosyntactic features and their evolution. Her work specifically contributes to the understanding of the relationship between individuals' cognitive biases at play during language learning and use and universal tendencies in morpheme order and paradigmatic morphological structure. She carries out her research within a comprehensive interdisciplinary framework combining methods from linguistic theory, quantitative typology and experimental linguistics.

Keynote Talk: Deep Phonology Features in Computational Phonology

Kyle Gorman
City University of New York
2023-07-14 13:00:00 –

Abstract: The linguist Ray Jackendoff considers "the discovery of distinctive features . . . to be a scientific achievement on the order of the discovery and verification of the periodic table in chemistry." Despite this, quite a bit of work in phonology—whether formal or computational—works with extensional sets of indivisible segments rather than the intensional, internally-structured definitions derived from distinctive features. In this talk I will first present philosophical and empirical arguments that phonological patterns are defined intensionally: segments are bundles of features and processes are defined in terms of "natural classes", or conjunctions of feature specifications. Then, I will argue against the received wisdom—both in formal and computational phonology—that phonological patterns should be specified "minimally", in terms of the fewest possible features consistent with the observed data. I show that feature minimization has undesirable cognitive and computational properties. In contrast, feature maximization—which, under the intensional view, is equivalent to set intersection—is empirically adequate and free of the problems that plague feature minimization.

Bio: Kyle Gorman is a professor of linguistics at the Graduate Center, City University of New York, and director of the master's program in computational linguistics. He is also a software engineer at Google LLC. Along with his collaborators, he is the author of Finite-State Text Processing and of award-winning papers at ACL 2019 and WNUT 6.

Table of Contents

Translating a low-resource language using GPT-3 and a human-readable dictionary
Micha Elsner and Jordan Needle 1

Evaluating Cross Lingual Transfer for Morphological Analysis: a Case Study of Indian Languages
Siddhesh Pawar, Pushpak Bhattacharyya and Partha Talukdar 14

Joint Learning Model for Low-Resource Agglutinative Language Morphological Tagging
Gulinigeer Abudouwaili, Kahaerjiang Abiderexiti, Nian Yi and Aishan Wumaier 27

Revisiting and Amending Central Kurdish Data on UniMorph 4.0
Sina Ahmadi and Aso Mahmudi 38

Investigating Phoneme Similarity with Artificially Accented Speech
Margot Masson and Julie Carson-berndsen 49

Generalized Glossing Guidelines: An Explicit, Human- and Machine-Readable, Item-and-Process Convention for Morphological Annotation
David R. Mortensen, Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja and Lori Levin 58

Jambu: A historical linguistic database for South Asian languages
Aryaman Arora, Adam Farris, Samopriya Basu and Suresh Kolichala 68

Lightweight morpheme labeling in context: Using structured linguistic representations to support linguistic analysis for the language documentation context
Bhargav Shandilya and Alexis Palmer 78

Improving Automated Prediction of English Lexical Blends Through the Use of Observable Linguistic Features
Jarem Saunders 93

Colexifications for Bootstrapping Cross-lingual Datasets: The Case of Phonology, Concreteness, and Affectiveness
Yiyi Chen and Johannes Bjerva 98

Character alignment methods for dialect-to-standard normalization
Yves Scherrer 110

SIGMORPHON–UniMorph 2023 Shared Task 0: Typologically Diverse Morphological Inflection
Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty and Ekaterina Vylomova 117

SIGMORPHON–UniMorph 2023 Shared Task 0, Part 2: Cognitively Plausible Morphophonological Generalization in Korean
Canaan Breiss and Jinyoung Jo 126

Morphological reinflection with weighted finite-state transducers
Alice Kwak, Michael Hammond and Cheyenne Wing 132

Linear Discriminative Learning: a competitive non-neural baseline for morphological inflection
Cheonkam Jeong, Dominic Schmitz, Akhilesh Kakolu Ramarao, Anna Stein and Kevin Tang 138

Tü-CL at SIGMORPHON 2023: Straight-Through Gradient Estimation for Hard Attention
Leander Girrbach 151

The BGU-MeLeL System for the SIGMORPHON 2023 Shared Task on Morphological Inflection
Gal Astrach and Yuval Pinter 166

Tü-CL at SIGMORPHON 2023: Straight-Through Gradient Estimation for Hard Attention
Leander Girrbach 171

Findings of the SIGMORPHON 2023 Shared Task on Interlinear Glossing
Michael Ginn, Sarah Moeller, Alexis Palmer, Anna Stacey, Garrett Nicolai, Mans Hulden and Miikka Silfverberg 186

LISN @ SIGMORPHON 2023 Shared Task on Interlinear Glossing
Shu Okabe and François Yvon 202

SigMoreFun Submission to the SIGMORPHON Shared Task on Interlinear Glossing
Taiqi He, Lindia Tjuatja, Nathaniel Robinson, Shinji Watanabe, David R. Mortensen, Graham Neubig and Lori Levin 209

An Ensembled Encoder-Decoder System for Interlinear Glossed Text
Edith Coates 217

Glossy Bytes: Neural Glossing using Subword Encoding
Ziggy Cross, Michelle Yun, Ananya Apparaju, Jata MacCabe, Garrett Nicolai and Miikka Silfverberg 222

The SIGMORPHON 2022 Shared Task on Cross-lingual and Low-Resource Grapheme-to-Phoneme Conversion
Arya D. McCarthy, Jackson L. Lee, Alexandra DeLucia, Travis Bartley, Milind Agarwal, Lucas F.E. Ashby, Luca Del Signore, Cameron Gibson, Reuben Raff and Winston Wu 230

SIGMORPHON 2022 Shared Task on Grapheme-to-Phoneme Conversion Submission Description: Sequence Labelling for G2P
Leander Girrbach 239

Low-resource grapheme-to-phoneme mapping with phonetically-conditioned transfer
Michael Hammond 245

A future for universal grapheme-phoneme transduction modeling with neuralized finite-state transducers
Chu-Cheng Lin Lin 249

Fine-tuning mSLAM for the SIGMORPHON 2022 Shared Task on Grapheme-to-Phoneme Conversion
Dan Garrette 250

Program

Friday, July 14, 2023

08:25 - 08:30 *Opening Remarks*

08:30 - 09:30 *Invited Talk:Carmen Saldana*

09:30 - 10:30 *Morning Session:Morphology*

Evaluating Cross Lingual Transfer for Morphological Analysis: a Case Study of Indian Languages
Siddhesh Pawar, Pushpak Bhattacharyya and Partha Talukdar

Joint Learning Model for Low-Resource Agglutinative Language Morphological Tagging
Gulinigeer Abudouwaili, Kahaerjiang Abiderexiti, Nian Yi and Aishan Wumaier

Generalized Glossing Guidelines: An Explicit, Human- and Machine-Readable, Item-and-Process Convention for Morphological Annotation
David R. Mortensen, Ela Gulsen, Taiqi He, Nathaniel Robinson, Jonathan Amith, Lindia Tjuatja and Lori Levin

Lightweight morpheme labeling in context: Using structured linguistic representations to support linguistic analysis for the language documentation context
Bhargav Shandilya and Alexis Palmer

10:30 - 11:00 *Morning Break*

11:00 - 12:00 *Post-break Session:Phonology and Phonetics*

Investigating Phoneme Similarity with Artificially Accented Speech
Margot Masson and Julie Carson-berndsen

Improving Automated Prediction of English Lexical Blends Through the Use of Observable Linguistic Features
Jarem Saunders

Colexifications for Bootstrapping Cross-lingual Datasets: The Case of Phonology, Concreteness, and Affectiveness
Yiyi Chen and Johannes Bjerva

Character alignment methods for dialect-to-standard normalization
Yves Scherrer

Friday, July 14, 2023 (continued)

12:00 - 13:00 *Lunch*

13:00 - 14:00 *Invited Talk:Kyle Gorman*

14:00 - 14:30 *ACL Findings Session*

AxomiyaBERTa:A Phonologically-aware Transformer Model for Assamese
Abhijnan Nath, Sheikh Mannan and Nikhil Krishnaswamy

Do Transformer Models do Phonology like a Linguist?
Saliha Muradoğlu and Mans Hulden

14:30 - 15:30 *Glossing Shared Task*

15:30 - 16:00 *Afternoon Break*

16:00 - 17:00 *Inflection Shared Task*

17:00 - 18:00 *Afternoon Session:Multilinguality and Language Resources*

Multilingual Sequence-to-Sequence Models for Hebrew NLP
Matan Eyal, Hila Noga, Roee Aharoni, Idan Szpektor and Reut Tsarfaty

Translating a low-resource language using GPT-3 and a human-readable dictionary
Micha Elsner and Jordan Needle

Revisiting and Amending Central Kurdish Data on UniMorph 4.0
Sina Ahmadi and Aso Mahmudi

Jambu: A historical linguistic database for South Asian languages
Aryaman Arora, Adam Farris, Samopriya Basu and Suresh Kolichala

Friday, July 14, 2023 (continued)