SIGMORPHON 2022

19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology

Proceedings of the Workshop

July 14, 2022

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 978-1-955917-82-7

Organizing Committee

Co-Chair

Garrett Nicolai, University of British Columbia
Eleanor Chodroff, University of York

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: Sandra Kübler, Indiana University
At Large: Çağrı Çöltekin, University of Tübingen

Program Committee

Reviewers

Khuyagbaatar Batsuren, National University of Mongolia
Canaan Breiss, MIT
Jane Chandlee, Haverford College
Çağrı Çöltekin, University of Tübingen
Daniel Dakota, Indiana University
Aniello De Santo, University of Utah
Ewan Dunbar, University of Toronto
Indranil Dutta, Jadavpur University
Micha Jacobs, University of Buffalo
Adam Jardine, Rutgers University
Greg Kobele, Universität Leipzig
Jordan Kodner, Stony Brook University
Sandra Kübler, Indiana University
Andrew Malouf, San Diego State University
Arya McCarthy, Johns Hopkins University
Kemal Oflazer, CMU Qatar
Gerald Penn, University of Toronto
Jelena Prokic, Universiteit Leiden
Jonathan Rawski, San Diego State University
Brian Roark, Google AI
Morgan Sonderegger, McGill University
Miikka Silfverberg, University of British Columbia
Kairit Sirts, University of Tarfu
Ekaterina Vylomova, University of Melbourne
Adam Wiemerslage, University of Colorado, Boulder
Adina Williams, Facebook AI Research
Colin Wilson, Johns Hopkins University
Anssi Yli-Jyrä, University of Helsinki
Changbing Yang, University of British Columbia

Keynote Talk: Parsing continuous speech into lexically bound phonetic sequences

Laura Gwilliams
University of California, San Francisco

Abstract: Speech consists of a continuously-varying acoustic signal. Yet human listeners experience it as sequences of discrete speech sounds, which are used to recognise words. To examine how the human brain appropriately sequences the speech signal, we recorded two-hour magnetoencephalograms from 21 subjects listening to short narratives. Our analyses show that the brain continuously encodes the three most recently heard speech sounds in parallel, and maintains this information long past the sensory input. Each speech sound has a representation that evolves over time, jointly encoding both its phonetic features and time elapsed since onset. This allows the brain to represent the relative order and phonetic content of the phonetic sequence. These dynamic representations are active earlier when phonemes are more predictable, and are sustained longer when lexical identity is uncertain. The flexibility in the dynamics of these representations paves the way for further understanding of how such sequences may be used to interface with higher order structure such as morphemes and words.

Bio: Laura Gwilliams received her PhD in Psychology with a focus in Cognitive Neuroscience from New York University in May 2020. Currently she is a post-doctoral researcher at UCSF, using MEG and ECoG data to understand how linguistic structures are parsed and composed while listening to continuous speech. The ultimate goal of Laura's research is to describe speech comprehension in terms of what operations are applied to the acoustic signal; which representational formats are generated and manipulated (e.g. phonetic, syllabic, morphological), and under what processing architecture.

Keynote Talk: Deep Phonology: Modeling language from raw acoustic data in a fully unsupervised manner

Gasper Begus
University of California, Berkeley

Abstract: In this talk, I propose that language and its acquisition can be modeled from raw speech data in a fully unsupervised manner with Generative Adversarial Networks (GANs) and that such modeling has implications both for the understanding of language acquisition and for the understanding of how deep neural networks learn internal representations. I propose a technique that allows us to "wug-test" neural networks trained on raw speech, analyze intermediate convolutional layers, and test a causal relationship between meaningful units in the output and latent/intermediate representations. I further propose an extension of the GAN architecture in which learning of meaningful linguistic units emerges from a requirement that the networks output informative data and includes both the perception and production principles. With this model, we can test what the networks can and cannot learn, how their biases match human learning biases in behavioral experiments, how speech processing in the brain compares to intermediate representations in deep neural networks (by comparing acoustic properties in intermediate convolutional layers and the brainstem), how symbolic-like rule-like computation emerges in internal representations, and what GAN's innovative outputs can teach us about productivity in human language. This talk also makes a more general case for probing deep neural networks with raw speech data, as dependencies in speech are often better understood than those in the visual domain and because behavioral data on speech (especially the production aspect) are relatively easily accessible.

Bio: Gašper Beguš an Assistant Professor at the Department of Linguistics at UC Berkeley where he directs the Berkeley Speech and Computation Lab. Before coming to Berkeley, he was an Assistant Professor at the University of Washington and before that he graduated with a Ph.D. from Harvard. His research focuses on developing deep learning models for speech data. More specifically, he trains models to learn representations of spoken words from raw audio inputs. He combines machine learning and statistical modeling with neuroimaging and behavioral experiments to better understand how neural networks learn internal representations in speech and how humans learn to speak.

Table of Contents

Program

Thursday, July 14, 2022

08:45 - 09:00 *Opening Remarks*

09:00 - 10:00 *Invited Talk 1: Laura Gwilliams: Parsing continuous speech into lexically bound phonetic sequences*

10:00 - 10:30 *Morning Break*

10:30 - 11:30 *Morning Session: Phonology and Phonetics*

Multidimensional acoustic variation in vowels across English dialects
James Tanner, Morgan Sonderegger and Jane Stuart-Smith

On Building Spoken Language Understanding Systems for Low Resourced Languages
Akshat Gupta

Domain-Informed Probing of wav2vec 2.0 Embeddings for Phonetic Features
Patrick Cormac English, John D. Kelleher and Julie Carson-Berndsen

Investigating phonological theories with crowd-sourced data: The Inventory Size Hypothesis in the light of Lingua Libre
Mathilde Hutin and Marc Allassonnière-Tang

11:30 - 12:30 *Lunch*

12:30 - 13:30 *Invited Talk 2: Gasper Begus: Deep Phonology: Modeling language from raw acoustic data in a fully unsupervised manner*

13:30 - 15:00 *Morning Session: Morphosyntax*

A Masked Segmental Language Model for Unsupervised Natural Language Segmentation
C.M. Downey, Fei Xia, Gina-Anne Levow and Shane Steinert-Threlkeld

Trees probe deeper than strings: an argument from allomorphy
Hossep Dolatian, Shiori Ikawa and Thomas Graf

Thursday, July 14, 2022 (continued)

Logical Transductions for the Typology of Ditransitive Prosody
Mai Ha Vu, Aniello De Santo and Hossep Dolatian

Subword-based Cross-lingual Transfer of Embeddings from Hindi to Marathi and Nepali
Niyata Bafna and Zdeněk Žabokrtský

Morphotactic Modeling in an Open-source Multi-dialectal Arabic Morphological Analyzer and Generator
Nizar Habash, Reham Marzouk, Christian Khairallah and Salam Khalifa

Unsupervised morphological segmentation in a language with reduplication
Simon Todd, Annie Huang, Jeremy Needle, Jennifer Hay and Jeanette King

15:00 - 15:30 *Afternoon Break*

15:30 - 17:45 *Shared Task Session*

17:45 - 18:00 *Closing Statements*