FieldMatters 2024

Field Matters. The Third Workshop on NLP Applications to Field Linguistics

Proceedings of the Workshop

August 16, 2024

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-158-2

Preface

Field Matters is a workshop focused on the various applications of NLP methods to field linguistics and the analysis of field data.

The primary pursuit of linguistic fieldwork is to document and describe languages. The former typically involves building a corpus and other resources for the language community, the latter ideally aims to produce a reference grammar. Advances in technology have enabled vast quantities of media to be recorded. These recordings (sound and/or video) require annotation and analysis for further linguistic research or resource development. This is often done manually. This processing bottleneck can be significantly sped up with computational methods.

NLP research focuses on developing methodology for different tasks that show significant performance in high-resource languages, allowing the automation of various routine tasks. The processing burdens faced by field linguists present a natural opportunity to marry NLP practices with the workflow of a field linguist. Similarly, the future development of NLP methods could gain from the linguistic diversity and unique tasks encountered during the description/documentation efforts.

With these in mind, Field Matters aims to provide a platform to deepen the dialogue between Computational and Field Linguists. Our workshop is hosted by the 62nd Annual Meeting of the Association for Computational Linguistics.

To highlight the highly interdisciplinary nature of our aim we invite field linguists and NLP researchers worldwide to our program committee. Each paper was assigned a field linguist along side minimally two computational linguists. Analyzing the difference in reviews of field linguists and NLP researchers, we have seen that reviewers provide different perspectives and give more diverse and fruitful feedback: while field linguists pay attention how practical this application could be or how well it fits in the idea of the workshop, NLP specialists comment on how relevant and accurate chosen methods are.

After the hard process of reviewing all submissions, the program committee chose nine papers for a poster or oral presentation at the workshop. Accepted papers illustrate the main idea of our workshop: how field linguistics may benefit from using contemporary methods of computational analysis and how the NLP community may evolve its methods with the help of under-resourced languages.

More specifically, chosen papers cover the following topics:

- Tools for fieldwork, including a language documentation tool and guidelines for human-computer interaction in the field of sociolinguistics;
- Creation of various corpora (both spoken and written);
- Speech and text processing tools for under-resourced languages and dialect variants;
- Phonology study with machine learning tools.

This year we have introduced the Special Track of Indigenous languages of Thaïland and South-East Asia in connection with co-location with ACL in Bangkok, Thailand.

We are incredibly grateful to the Field Matters program committee, who worked on peer review to give meaningful comments for each submission and made this workshop possible. We want to thank the invited speakers, Emily Prud'hommeaux, Genta Indra Winata, and Alham Fikri Aji, for contributing to the program. We would also like to mention all the authors who submitted their papers to our workshop, and we hope to continue to serve as a link between NLP specialists and field linguists.

Organizing Committee

General Chairs

Oleg Serikov, King Abdullah University of Science and Technology (KAUST)
Ekaterina Voloshina, University of Gothenburg, Chalmers University of Technology
Anna Postnikova
Saliha Muradoğlu, The Australian National University (ANU)
Eric Le Ferrand, Boston College
Elena Klyachko
Ekaterina Vylomova, University of Melbourne
Tatiana Shavrina, Meta
Francis Tyers, Indiana University

Program Committee

Program Committee

I Wayan Arka, Australian National University
Timofey Arkhangelskiy, Universität Hamburg
Alexandre Arkhipov, Universität Hamburg and Lomonosov Moscow State University
James Bednall, Charles Darwin University
Anton Buzanov, HSE University
Shobhana Lakshmi Chelliah, Indiana University at Bloomington
Michael Daniel, Collegium de Lyon
Don Daniels, University of Oregon
Kilian Evang, Heinrich Heine University Düsseldorf
Junior Pierre Eden Fevrier, University at Buffalo
Konstantin V. Filatov, HSE University
James Gray, The Australian National University (ANU)
Harald Hammarström, Uppsala University
Huade Huang, The Australian National University (ANU)
Elena Klyachko
Ezequiel Koile, Max Planck Institute for the Science of Human History
Zoey Liu, University of Florida
Tessa Masis, University of Massachusetts at Amherst
Vladislav Mikhailov, University of Oslo
David R Mortensen, Carnegie Mellon University
Saliha Muradoglu, The Australian National University (ANU)
Bruno Olsson, Universität Regensburg
Michael Proctor, Macquarie University
Emily Prud'hommeaux, Boston College
Oleg Serikov, King Abdullah University of Science and Technology
Tatiana Shavrina, Meta
Nick Thieberger, University of Melbourne
José Carlos Antonio Pérez Vargas, University at Buffalo
Albert Ventayol-Boada, University of California, Santa Barbara
Ekaterina Voloshina, Göteborg University and Chalmers University of Technology
Sasha Wilmoth, University of Melbourne
He Zhou, The Hong Kong Polytechnic University

Table of Contents

Program

Friday, August 16, 2024

09:00 - 09:30 *Opening word*

09:30 - 10:30 *Invited talk. Emily Prud'hommeaux*

10:30 - 11:00 *Coffee break*

11:00 - 11:10 *Introduction to Special Track: Indigenous languages of Thailand and South-East Asia*

11:10 - 12:10 *Talks*

Leveraging Deep Learning to Shed Light on Tones of an Endangered Language: A Case Study of Moklen
Sireemas Maspong, Francesco Burroni, Teerawee Sukanchanon, Warunsiri Pornpottanamas and Pittayawat Pittayaporn

Documenting Endangered Languages with LangDoc: A Wordlist-Based System and A Case Study on Moklen
Piyapath T Spencer

Zero-shot Cross-lingual POS Tagging for Filipino
Jimson Paulo Layacan, Isaiah Edri W. Flores, Katrina Bernice M. Tan, Ma. Regina E. Estuar, Jann Railey Montalan and Marlene M. De Leon

12:10 - 12:20 *Break*

12:20 - 13:20 *Invited talk. Genta Winata and Alham Fikri Aji*

13:20 - 13:30 *Break*

13:30 - 15:30 *Talks*

The Parallel Corpus of Russian and Ruska Romani Languages
Kirill Koncha, Abina Abina, Kazakova Tatiana and Gloria Rozovskaya

ManWav: The First Manchu ASR Model
Jean Seo, Minha Kang, SungJoo Byun and Sangah Lee

Friday, August 16, 2024 (continued)

User-Centered Design of Digital Tools for Sociolinguistic Studies in Under-Resourced Languages
Jonas Adler, Carsten Scholle, Daniel Buschek, Nicolo' Brandizzi and Muhadj Adnan

A Comparative Analysis of Speaker Diarization Models: Creating a Dataset for German Dialectal Speech
Lea Fischbach

Noise Be Gone: Does Speech Enhancement Distort Linguistic Nuances?
Iñigo Parra

Comparing Kaldi-Based Pipeline Elpis and Whisper for Čakavian Transcription
Austin Jones, Shulin Zhang, John T. Hale, Margaret Renwick, Zvjezdana Vrzic and Keith Langston

15:30 - 16:00 *Coffee break*

16:00 - 17:30 *Plenar discussion*