FieldMatters 2023

The Second Workshop on NLP Applications to Field Linguistics (Field Matters)

Proceedings of the Workshop

May 6, 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 978-1-959429-60-9

Preface by the General Chair

Field Matters is a workshop focused on various applications of NLP methods to field linguistics and analysis of field data with the help of computational linguistics.
On the one hand, field linguists document language data, but the fieldwork involves tons of manual annotation or analysis, which might be significantly sped up with computational instruments. On the other hand, NLP research brought methods for different tasks that show significant performance in high-resource languages, allowing to automate various routine tasks. The future development of NLP methods could gain from the language diversity of under-resourced languages. Field Matters is aimed to combine linguistic fieldwork and NLP methods. Our workshop is hosted by the 17th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2023).
To provide the comprehensive diverse expertise in a multidisciplinary setting, for the second time we invite linguists and NLP researchers worldwide to our program committee. 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 underresourced languages.
More specifically, chosen papers cover the following topics:

- use of ASR into the field linguistic pipeline
- use of computational methods to provide deterministic grounding for the language documentation insights
- incorporating linguistic knowledge to the neural language processing algorithms despite the low-resourced setting
- using Information Extraction algorithms to support the language documentation
- building tools for native speakers community

Following the key insight of the FM2022, in some studies, the collaborative nature of the process has taken its place, making the results useful for both researchers and native speakers.
Notably, the recently popularized Limitations section has proven itself useful. Several papers contain meaningful insights into the state of the field or language nuanced details worth attention themselves.
Given 24 submissions in total (including 3 papers submitted through the ACL Findings program), the acceptance rate is $11/24$, with 4 papers selected for oral presentation.
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, Lane Swartz and Emmanuel Schang, 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, Independent Researcher
Ekaterina Voloshina, Independent Researcher
Anna Postnikova, Independent Researcher
Elena Klyachko, Independent Researcher
Ekaterina Vylomova, University of Melbourne
Tatiana Shavrina, Independent Researcher
Eric Le Ferrand, Universite Grenoble Alpes, Universite d'Orleans
Valentin Malykh, Huawei
Francis Tyers, Indiana University, HSE University
Timofey Arkhangelskiy, University of Hamburg
Vladislav Mikhailov, Independent Researcher

Program Committee

Chairs

Elena Klyachko, HSE, Institute of Linguistics RAS
Oleg Serikov, DeepPavlov, AIR Institute, HSE University
Ekaterina Voloshina, AIRI

Program Committee

Dmitry Abulkhanov, Huawei Noah's Ark
Alexandre Arkhipov, Universität Hamburg
Harald Hammarström, Uppsala University
Ezequiel Koile, Max Planck Institute for Evolutionary Anthropology
Éric Le Ferrand, Université d'Orléans
Zoey Liu, Department of Linguistics, University of Florida
Valentin Malykh, Huawei Noah's Ark Lab / Kazan Federal University
Tessa Masis, University of Massachusetts Amherst
Vladislav Mikhailov, HSE University
Saliha Muradoglu, The Australian National University
Vitaly Protasov, AIRI
Emily Prud'hommeaux, Boston College
Tatiana Shavrina, AIRI
He Zhou, Indiana University
Nadezhda Zueva, VK

Table of Contents

Program

Friday, May 5, 2023

09:00 - 10:30 *Invited talk. Lane Schwartz.*

11:15 - 12:45 *Invited talk. Emmanuel Schang.*

12:45 - 14:15 *lunch break*

14:15 - 15:45 *Presentations*

Speech Database (Speech-DB) – An on-line platform for storing, validating, searching, and recording spoken language data
Jolene Poulin, Daniel Dacanay and Antti Arppe

Improving Low-resource RRG Parsing with Structured Gloss Embeddings
Roland Eibers, Kilian Evang and Laura Kallmeyer

A Quest for Paradigm Coverage: The Story of Nen
Saliha Muradoglu, Hanna Suominen and Nicholas Evans

Unsupervised part-of-speech induction for language description: Modeling documentation materials in Kolyma Yukaghir
Albert Ventayol-boada, Nathan Roll and Simon Todd

15:45 - 16:30 *Coffee Break*

16:30 - 18:00 *Presentations*

Automated speech recognition of Indonesian-English language lessons on YouTube using transfer learning
Zara Maxwelll-smith and Ben Foley

Application of Speech Processes for the Documentation of Kréyòl Gwadloupéyen
Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux and Emmanuel Schang

ASR pipeline for low-resourced languages: A case study on Pomak
Chara Tsoukala, Kosmas Kritsis, Ioannis Douros, Athanasios Katsamanis, Nikolaos Kokkas, Vasileios Arampatzakis, Vasileios Sevetlidis, Stella Markantonatou and George Pavlidis

Friday, May 5, 2023 (continued)

Approaches to Corpus Creation for Low-Resource Language Technology: the Case of Southern Kurdish and Laki
Sina Ahmadi, Zahra Azin, Sara Belelli and Antonios Anastasopoulos

AraDiaWER: An Explainable Metric For Dialectical Arabic ASR
Abdulwahab Sahyoun and Shady Shehata

Multilingual Automatic Extraction of Linguistic Data from Grammars
Albert Kornilov