Field Matters

Proceedings of the Fifth Workshop on NLP Applications to Field Linguistics

March 29, 2026

ISBN 979-8-89176-376-0

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 19th Conference of the European Chapter of the Association for Computational Linguistics in Rabat, Morocco.

Field Matters 2026 continued to provide field linguists expert reviews, a distinct feature of the review process introduced two years ago. Each paper was assigned a field linguist alongside 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 to 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 seven 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.

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 speaker, Alexis Palmer, Associate Professor at the University of Colorado Boulder. We would also like to acknowledge 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

Éric Le Ferrand (Boston College)

Elena Klyachko

Shu Okabe (Technische Universität München)

Ekaterina Voloshina (Chalmers University of Technology, University of Gothenburg)

Oleg Serikov (Palisade Research)

Tatiana Shavrina (Meta)

Ekaterina Vylomova (University of Melbourne)

Table of Contents

Conference Program

Automated Quality Control for Language Documentation: Detecting Phonotactic Inconsistencies in a Kokborok Wordlist
Kellen Parker van Dam and Abishek Stephen

What NLP Gets Wrong About Contact: Implications for Field Linguistic Evidence
Manodyna K H

Hybrid Neural-LLM Pipeline for Morphological Glossing in Endangered Language Documentation: A Case Study of Jungar Tuvan
Siyu Liang, Talant Mawkanuli and Gina-Anne Levow

Linguistically Informed Tokenization Improves ASR for Underresourced Languages
Massimo Marie Daul, Alessio Tosolini and Claire Bowern

Short-form verbal arts as a speech data resource in the field
Matthew Faytak, Tianle Yang, Pius Wuchu Akumbu, Ivo Forghema Njuasi and Éric Le Ferrand

Quantitative Lect Description: A Case Study of Lemko from the Field Data of 1920s-1930s
Ilia Afanasev

The order of subject, object and verb in Tatyshly Udmurt
Daria Belova and Irina Khomchenkova