FieldMatters 2025

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

Proceedings of the Workshop

August 1, 2025

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 979-8-89176-282-4

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 63rd Annual Meeting of the Association for Computational Linguistics in Vienna, Austria.

Field Matters 2025 continued to provide field linguists expert reviews, a distinct feature of the review process introduced one year 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.

This year, *Field Matters* shared a venue with *SigTyp*, a workshop dedicated to linguistic typology and multilingual NLP. Although the ultimate goals of *Field Matters* and *SigTyp* differ, the co-location provided a valuable opportunity for both communities to learn from one another. Careful consideration suggested we share our space while keeping the publication processes separate. This gave us twice as many keynotes and a tightly packed schedule of oral presentations. We anticipate twice as fruitful discussions in the hallways, though the dual load brings an intense workload for both organizers and participants of the one-day event, reflecting the growing audience of both workshops.

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:

- The creation of datasets and tools for field linguistics
- ASR and speech processing to address the transcription bottleneck
- Machine translation for very low resource languages

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, Alexis Michaud, researcher at LACITO-CNRS in Paris, France, and Eduardo Sanchez, research scientist at Meta. 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.

Keynote Talks

Alexis Michaud, LACITO CNRS
“Archives, Algorithms, and Alliances: Grounding NLP in the Realities of Language Documentation”

This talk offers a linguist’s perspective on the evolving place of NLP in language documentation, focusing on the interplay between archives (as both legacy and infrastructure), algorithms (with ASR on the Na language as an example), and alliances (the human networks that sustain such work). Drawing on experience within “Computational Language Documentation” projects led by computer scientists, I reflect on shared goals, realistic expectations, and the practical conditions required to keep interdisciplinary teams motivated over time.

Eduardo Sanchez, Meta
“A few good texts: how small sets of high quality linguistic data power massive multilinguality in language models”

While scale remains a key driver of performance in multilingual language models, it’s not always an option, especially for low-resource languages where data is scarce or noisy. We’ll explore how small, high-quality datasets can play a surprisingly powerful role in enabling multilinguality, especially where coverage gaps exist. Beyond parallel corpora, we’ll show how strategic use of linguistic resources can complement large-scale training, improve generalization, and unlock better performance for underserved languages. A few good texts, chosen well, may be worth billions of tokens, and for many languages, they may be the key to ensuring visibility, usability, and survival in the digital age.

Organizing Committee

General Chairs

Éric Le Ferrand, Boston College
Elena Klyachko
Anna Postnikova
Oleg Serikov
Tatiana Shavrina, Meta
Ekaterina Voloshina, University of Gothenburg, Chalmers University of Technology
Ekaterina Vylomova, University of Melbourne

Program Committee

Program Chairs

Eric Le Ferrand, Boston College
Elena Klyachko
Anna Postnikova
Oleg Serikov, King Abdullah University of Science and Technology
Tatiana Shavrina, Meta
Ekaterina Voloshina, Göteborg University and Chalmers University of Technology
Ekaterina Vylomova, The University of Melbourne

Reviewers

Angelina Aspra Aquino, Alexandre Arkhipov

James Bednall, Anton Buzanov

Michael Daniel

Harald Hammarström, William N. Havard

Elena Klyachko, Ezequiel Koile

Jordan Lachler, Eric Le Ferrand, Kate L Lindsey

Tessa Masis, Field Matters, Saliha Muradoglu

Shu Okabe

Anna Postnikova, Michael Proctor

Emmanuel Schang, Oleg Serikov, Tatiana Shavrina

Nick Thieberger

Alexey Vinyar, Ekaterina Voloshina, Ekaterina Vylomova

Table of Contents