

SIGTYP 2025

**The 7th Workshop on Research in Computational Linguistic  
Typology and Multilingual NLP**

**Proceedings of the Workshop**

August 1, 2025

©2025 Association for Computational Linguistics

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](mailto:acl@aclweb.org)

ISBN 979-8-89176-281-7

# Introduction

We are pleased to present the proceedings of SIGTYP 2025, the seventh edition of the Workshop on Research in Computational Linguistic Typology and Multilingual Natural Language Processing. This year, the workshop is held as a joint event with FieldMatters and is co-located with the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025), taking place in Vienna, Austria. Building on the success of previous editions from 2019 through 2024, SIGTYP continues to serve as a platform for fostering dialogue between the fields of linguistic typology and multilingual NLP. Our core mission remains the same: to raise awareness of typological diversity and to explore how insights from linguistic typology can inform, enrich, and challenge computational methods in cross-lingual and multilingual settings. We are particularly committed to the development of truly inclusive NLP methods that serve a broad and typologically diverse range of languages.

SIGTYP 2025 invites contributions at the intersection of typology and NLP, with key areas of focus including:

- The integration of typological features in multilingual learning and language transfer;
- The development of unified linguistic taxonomies and cross-lingual resources;
- Automatic inference of typological features using machine learning;
- Enhancing interpretability of multilingual models through typological knowledge;
- Collaborative approaches to improving typological databases;
- Addressing the challenges of cross-lingual annotation and defining linguistic universals;
- Language-specific studies aimed at supporting or revising typological claims.

This year's program includes 2 keynote talks, 15 archival papers and 2 extended abstracts. We are honored to host Robert Forkel and Lisa Bylinina as invited speakers, whose work exemplifies the interdisciplinary spirit of the workshop. We extend our sincere thanks to all authors for their high-quality submissions, to the program committee for their diligent and insightful reviews, and to all participants who contribute to the vibrancy and impact of SIGTYP. For more information, including proceedings and shared task resources, please visit the workshop website: [website: https://sigtyp.github.io/ws2025-sigtyp.html](https://sigtyp.github.io/ws2025-sigtyp.html)

# Organizing Committee

## Organizing Committee

Michael Hahn, Saarland University  
Priya Rani, University of Galway  
Ritesh Kumar, Dr. Bhimrao Ambedkar University  
Andreas Shcherbakov, The University of Melbourne  
Alexey Sorokin, Yandex and Lomonosov Moscow State University  
Oleg Serikov, King Abdullah University of Science and Technology  
Ryan Cotterell, Swiss Federal Institute of Technology  
Ekaterina Vylomova, The University of Melbourne

# Program Committee

## Program Chairs

Michael Hahn, Saarland University  
Priya Rani, University of Galway  
Oleg Serikov, King Abdullah University of Science and Technology  
Andreas Shcherbakov, University of Melbourne  
Ritesh Kumar, Dr. Bhimrao Ambedkar University  
Alexey Sorokin, Yandex and Lomonosov Moscow State University  
Ekaterina Vylomova, The University of Melbourne  
Ryan Cotterell, Swiss Federal Institute of Technology

## Reviewers

Emily Ahn  
  
Barend Beekhuizen, Claire Bower  
  
Giuseppe G. A. Celano, Ryan Cotterell, Jannic Alexander Cutura  
  
Rena Wei Gao  
  
Michael Hahn, Borja Herce  
  
Elisabetta Jezek  
  
Ritesh Kumar, Kemal Kurniawan  
  
M. Dolores Jiménez López  
  
Aso Mahmudi, Raphael Merx  
  
Gaurav Negi  
  
Devishree Pillai, Edoardo Ponti  
  
Priya Rani  
  
Oleg Serikov, Andreas Shcherbakov, Alexey Sorokin, Richard Sproat  
  
Ekaterina Vylomova  
  
Jinrui Yang

**Keynote Talk**  
**Connecting the dots - growing an eco-system for  
cross-linguistic data**

**Robert Forkel**

Max Planck Institute for Evolutionary Anthropology

**2025-08-01 – Room: TBD**

**Abstract:** One of the key contributions typology can make to multilingual NLP is a fuller picture of the diversity of the world's languages. This diversity is also reflected in widely varying documentation across languages. Thus, informing computational approaches to language processing by this diversity requires operationalizing a variety of data types describing very different languages. Getting a computational grasp on cross-linguistic information has been the main motivation behind CLDF - the Cross-Linguistic Data Formats. This talk will explore the eco-system of cross-linguistic data that is now opened up via CLDF.

**Bio:** Robert Forkel leads the research data management group and serves as a scientific programmer in the Department of Linguistic and Cultural Evolution at the Max Planck Institute for Evolutionary Anthropology in Germany. His current work centers on developing software solutions to collect, curate, and publish large-scale databases for linguistic and cultural research. He is also interested in the role of data in scientific research, with a particular focus on reproducibility. In addition, he contributes to open-source software packages such as LingPy.

# Keynote Talk

## (L)LMs and language theory

**Lisa Bylinina**

Utrecht University

**2025-08-01 – Room: TBD**

**Abstract:** One of the central questions in linguistic typology is: What constrains the space of natural languages? In a somewhat narrower formulation: How do different grammatical properties of a language relate to each other, and why are some combinations of features that would, in principle, be possible, in fact not attested? I would like to put these questions in the context of recent language models. Can (L)LMs help us understand interconnections within linguistic grammatical systems? I will argue for a moderately optimistic view and suggest some ways to make progress in this direction, with a focus on the linguistic generalisations (L)LMs make under different training conditions. My goal is to encourage discussion about the usefulness of (L)LMs for theoretical and typological linguistic research.

**Bio:** Lisa Bylinina is an Assistant Professor of Computational Linguistics (UD1) at Utrecht University, where she is part of the Language and Communication group within the Institute for Language Sciences. She is also an active member of the NLP@U special interest group. Her research interests lie at the intersection of theoretical linguistics and natural language processing. Before joining Utrecht University in September 2024, she held the position of Assistant Professor at the University of Groningen, in the Computational Linguistics Group at the Center for Language and Cognition (CLCG). At Utrecht, she teaches in the Applied Data Science master's program and the bachelor's program in Communication and Information Science. She is open to supervising (research) master's theses in data science, artificial intelligence, and theoretical linguistics, particularly in semantics.

## Table of Contents

|                                                                                                                                                                                                                                                                 |     |
|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| <i>InstructionCP: A Simple yet Effective Approach for Transferring Large Language Models to Target Languages</i>                                                                                                                                                |     |
| Kuang-Ming Chen, Jenq-Neng Hwang and Hung-yi Lee . . . . .                                                                                                                                                                                                      | 1   |
| <i>Analyzing the Linguistic Priors of Language Models with Synthetic Languages</i>                                                                                                                                                                              |     |
| Alessio Tosolini and Terra Blevins . . . . .                                                                                                                                                                                                                    | 7   |
| <i>Unstable Grounds for Beautiful Trees? Testing the Robustness of Concept Translations in the Compilation of Multilingual Wordlists</i>                                                                                                                        |     |
| David Snee, Luca Ciucci, Arne Rubehn, Kellen Parker Van Dam and Johann-Mattis List . . . . .                                                                                                                                                                    | 16  |
| <i>Annotating and Inferring Compositional Structures in Numeral Systems Across Languages</i>                                                                                                                                                                    |     |
| Arne Rubehn, Christoph Rzymiski, Luca Ciucci, Katja Bocklage, Alžběta Kučerová, David Snee, Abishek Stephen, Kellen Parker Van Dam and Johann-Mattis List . . . . .                                                                                             | 29  |
| <i>Beyond the Data: The Impact of Annotation Inconsistencies in UD Treebanks on Typological Universals and Complexity Assessment</i>                                                                                                                            |     |
| Antoni Brosa Rodríguez and M. Dolores Jiménez López . . . . .                                                                                                                                                                                                   | 43  |
| <i>Beyond cognacy</i>                                                                                                                                                                                                                                           |     |
| Gerhard Jäger . . . . .                                                                                                                                                                                                                                         | 52  |
| <i>SenWiCh: Sense-Annotation of Low-Resource Languages for WiC using Hybrid Methods</i>                                                                                                                                                                         |     |
| Roksana Goworek, Harpal Singh Karicut, Hamza Shezad, Nijaguna Darshana, Abhishek Mane, Syam Bondada, Raghav Sikka, Ulvi Mammadov, Rauf Allahverdiyev, Sriram Satkirti Purighella, Paridhi Gupta, Muhinyia Ndegwa, Bao Khanh Tran and Haim Dubossarsky . . . . . | 61  |
| <i>XCOMPS: A Multilingual Benchmark of Conceptual Minimal Pairs</i>                                                                                                                                                                                             |     |
| Linyang He, Ercong Nie, Sukru Samet Dindar, Arsalan Firoozi, Van Nguyen, Corentin Puffay, Riki Shimizu, Haotian Ye, Jonathan Brennan, Helmut Schmid, Hinrich Schuetze and Nima Mesgarani . . . . .                                                              | 75  |
| <i>Tone in Perspective: A Computational Typological Analysis of Tone Function in ASR</i>                                                                                                                                                                        |     |
| Siyu Liang and Gina-Anne Levow . . . . .                                                                                                                                                                                                                        | 82  |
| <i>A discovery procedure for synlexification patterns in the world's languages</i>                                                                                                                                                                              |     |
| Hannah S. Rognan and Barend Beekhuizen . . . . .                                                                                                                                                                                                                | 93  |
| <i>Construction-Based Reduction of Translationese for Low-Resource Languages: A Pilot Study on Bavarian</i>                                                                                                                                                     |     |
| Peiqin Lin, Marion Thaler, daniela.goschala@campus.lmu.de daniela.goschala@campus.lmu.de, Amir Hossein Kargaran, Yihong Liu, Andre Martins and Hinrich Schuetze . . . . .                                                                                       | 114 |
| <i>High-Dimensional Interlingual Representations of Large Language Models</i>                                                                                                                                                                                   |     |
| Bryan Wilie, Samuel Cahyawijaya, Junxian He and Pascale Fung . . . . .                                                                                                                                                                                          | 122 |
| <i>Domain Meets Typology: Predicting Verb-Final Order from Universal Dependencies for Financial and Blockchain NLP</i>                                                                                                                                          |     |
| Zichao Li and Zong Ke . . . . .                                                                                                                                                                                                                                 | 156 |
| <i>Token-level semantic typology without a massively parallel corpus</i>                                                                                                                                                                                        |     |
| Barend Beekhuizen . . . . .                                                                                                                                                                                                                                     | 165 |
| <i>Are Translated Texts Useful for Gradient Word Order Extraction?F</i>                                                                                                                                                                                         |     |
| Amanda Kann . . . . .                                                                                                                                                                                                                                           | 177 |