TLT 2024

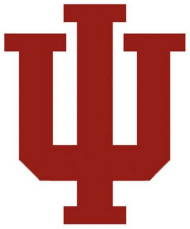
The 22nd Workshop on Treebanks and Linguistic Theories (TLT 2024)

Proceedings of the Conference

December 5-6, 2024

The TLT organizers gratefully acknowledge the support from the following sponsors.

University of Hamburg, Indiana University, supported by the DAAD in the program University Partnerships with Eastern Europe with funds from the Federal Foreign Office of Germany, and the SFB 1102



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 None

Introduction

The 22nd International Workshop on Treebanks and Linguistic Theories (TLT 2024) follows an annual series that started in 2002 in Sozopol, Bulgaria. TLT addresses all aspects of treebank design, development, and use. "Treebank" is taken in a broad sense, comprising any spoken, signed, or written data augmented with computationally processable annotations of linguistic structure at various levels.

For the first time, TLT is being hosted by the Hamburg KorpusLab on December 5-6, 2024 in Hamburg, Germany. The KorpusLab is a research group led by Heike Zinsmeister at the Institute for German Language and Literature (Institut für Germanistik) at the University of Hamburg. Information about the group's research projects and other activities are collected on the KorpusLab website.

From the papers submitted to TLT 2024, we accepted 8 archival submssions as well as 1 non-archival submission. The papers range in topics from UD treebanks for new languages to coreference and information status annotations on top of UD annotations for the literary domain, and a novel approach to parsing dependencies using shallow information. For the first time, TLT offered the option of non-archival submissions.

When the call for the workshop was published, a member of the community expressed concerns about the relevance of linguistically annotated resources in the area of large language models (LLMs) and questioned the appropriateness of continuing research on creating and analyzing such resources. Addressing this concern, we organized a panel discussion on "Treebanks and linguistic annotation in the area of LLMs".

Our sincere thanks go to everyone who is making this event possible: everybody who submitted papers; all the reviewers; our invited speakers and panelists, and the SFB 1102 at Saarland University for funding an invited speaker.

Finally, we would also like to thank ACL SIGPARSE for its endorsement and the ACL Anthology for publishing the proceedings.

Daniel Dakota, Sarah Jablotschkin, Sandra Kübler, Heike Zinsmeister (TLT2024 Chairs)
December 2024

Organizing Committee

TLT2024 Chairs

Daniel Dakota, Indiana University
Sarah Jablotschkin, University of Hamburg
Sandra Kübler, Indiana University
Heike Zinsmeister, University of Hamburg

Program Committee

Reviewers

Ann Bies, University of Pennsylvania
Gosse Bouma, University of Groningen
Miriam Butt, Universität Konstanz
Éric Villemonte de la Clergerie, INRIA
Eva Hajicova, Charles University Prague
Lori Levin, Carnegie Mellon University
Wolfgang Menzel, University of Hamburg
Adam Meyers, New York University
Jiří Mírovský, Charles University Prague
Kaili Müürisep, University of Tartu
Joakim Nivre, Uppsala University
Petya Osenova, Bulgarian Academy of Sciences
Daniel Zeman, Charles University Prague

Keynote Talk: Multilingual Coreference and Treebanking: Benefits of Interaction

Anna Nedoluzhko
Charles University, Prague

Abstract: Several years ago, we created CorefUD, a harmonized collection of coreference datasets for multiple languages. This collection has grown steadily, with new languages and datasets added each year. Currently, CorefUD 1.2 includes 21 datasets across 15 languages. CorefUD is compatible with morphosyntactic annotations in the Universal Dependencies (UD) framework, highlighting the close relationship between two types of linguistic annotation: coreference and syntax. But how do these annotations interact? Do UD tree structures correspond to mention spans in coreference annotations? Are syntactic heads in UD equivalent to the head mentions in coreference annotation? Can reconstructed empty nodes in enhanced UD effectively align with zero anaphora? And how do zeros in coreference relate to syntactic structures across the diverse languages in the collection? In the talk, I will address these questions with a specific focus on zero anaphora which was the special topic of the recent CRAC shared task on multilingual coreference resolution.

Keynote Talk: Increasing Language Diversity in NLP

Marcel Bollmann
Linköping University

Abstract: Linguistic diversity in NLP remains an important challenge, with many languages lagging behind in terms of available data and resources for training and evaluation of NLP models. In this talk, I will present CreoleVal, a project aimed at providing an evaluation benchmark for several Creole languages. I will discuss why we chose to work on Creoles in particular, what kinds of data and annotations we produced for CreoleVal, and what challenges we encountered in the process. Finally, I will give an outlook on challenges around data and data annotation in the TrustLLM project, an ongoing EU-funded project on creating trustworthy LLMs for the Germanic languages.

Treebanks and Linguistic Annotation in the Area of LLMs

The panel discussed the impact Large Language Models (LLMs) have had on the current state of treebank design and development, as well as their continued impact on the future of the field. Topics considered included:

- Do LLMs make treebanks redundant?
- What can we learn from treebanks that we can't learn from LLMs?
- Is it still justified to spend money on creating and maintaining treebanks?

Invited Panel Members

Marcel Bollmann, Linköping University
Daniel Dakota, Indiana University
Anna Nedoluzhko, Charles University Prague
Sandra Kübler, Indiana University
Juri Opitz, University of Zurich

Non-Archival Abstracts

UD for German Poetry
Stefanie Dipper and Ronja Laarmann-Quant
Ruhr-Universtät Bochum

This article deals with the syntactic analysis of German-language poetry from different centuries. We use Universal Dependencies (UD) as our syntactic framework. We discuss particular challenges of the poems in terms of tokenization, sentence boundary recognition and special syntactic constructions. Our annotated pilot corpus currently consists of 20 poems with a total of 2,162 tokens, which originate from the PoeTree.de corpus. We present some statistics on our annotations and also evaluate the automatic UD annotation from PoeTree.de using our annotations.

Table of Contents