

TLT 2025

**23rd International Workshop on Treebanks and Linguistic
Theories (TLT, SyntaxFest 2025)**

Proceedings

August 28-29, 2025

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

VITASIS



Mestna občina
Ljubljana



Flanders
State of the Art



Organized by



As part of SyntaxFest 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

ISBN 979-8-89176-291-6

Introduction

The 23rd International Workshop on Treebanks and Linguistics (TLT) 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. This year, TLT took place at SyntaxFest 2025 in Ljubljana, Slovenia, which brought together five related but independent events:

- 18th International Conference on Parsing Technologies (IWPT 2025)
- 8th Universal Dependencies Workshop (UDW 2025)
- 8th International Conference on Dependency Linguistics (DepLing 2025)
- 23rd Workshop on Treebanks and Linguistic Theories (TLT 2025)
- 3rd Workshop on Quantitative Syntax (QUASY 2025)

In addition, a pre-conference workshop organized by the COST Action CA21167 – Universality, Diversity and Idiosyncrasy in Language Technology (UniDive) was held prior to the main event, with dedicated sessions on the 1st UniDive Shared Task on Morphosyntactic Parsing and the 2nd Workshop on Universal Dependencies for Turkic Languages.

SyntaxFest 2025 continues the tradition of SyntaxFest 2019 (Paris, France), SyntaxFest 2021 (Sofia, Bulgaria), and GURT/SyntaxFest 2023 (Washington DC, USA) in bringing together multiple events that share a common interest in using corpora and treebanks for empirically validating syntactic theories, studying syntax from quantitative and theoretical points of view, and training machine learning models for natural language processing. Much of this research is increasingly multilingual and cross-lingual and requires continued systematic analysis from various theoretical, applied, and practical perspectives. By co-locating these workshops under a shared umbrella, SyntaxFest fosters dialogue between overlapping research communities and supports innovation at the intersection of linguistics and language technology. As in previous editions, all five workshops at SyntaxFest 2025 shared a common submission and reviewing process, with a unified timeline, identical submission formats, and a shared program committee. During submission, authors could indicate one or more preferred venues, but the final assignment of papers was determined by the collective program chairs, composed of the individual workshop chairs, based on thematic alignment. All accepted submissions were peer-reviewed by at least three reviewers from the shared program committee.

In total, SyntaxFest 2025 received 94 submissions, of which 73 (78%) were accepted for presentation. The final program included a total of 47 long papers, 21 short papers, and 5 non-archival contributions, distributed across the five workshops: 5 papers were presented at IWPT (2 long, 3 short); 20 at UDW (14 long, 5 short, 1 non-archival); 16 at DepLing (12 long, 2 short, 2 non-archival); 18 at TLT (10 long, 7 short, 1 non-archival); and 14 at QUASY (9 long, 4 short, 1 non-archival).

Our sincere thanks go to everyone who made this event possible. We thank all authors for their submissions and the reviewers for their time and thoughtful feedback, which contributed to a diverse and high-quality program. Special thanks go to the local organizing team at the University of Ljubljana and the Slovene Language Technologies Society for hosting the event, and to the sponsors for their generous support. Finally, we gratefully acknowledge ACL SIGPARSE for endorsing the event and the ACL Anthology for publishing the proceedings.

Kenji Sagae, Stephan Oepen (IWPT 2025 Chairs)

Gosse Bomma, Çağrı Çöltekin (UDW 2025 Chairs)

Eva Hajičová, Sylvain Kahane (DepLing 2025 Chairs)

Heike Zinsmeister, Sarah Jablotschkin, Sandra Kübler (TLT 2025 Chairs)

Xinying Chen, Yaqin Wang (QUASY 2025 Chairs)
Kaja Dobrovoljc (SyntaxFest 2025 Organization Chair)

Ljubljana, August 2025

Organizing Committee

TLT Chairs

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

DepLing Chairs

Eva Hajičová, Charles University, Prague
Sylvain Kahane, Université Paris Nanterre

UDW Chairs

Gosse Bomma, University of Groningen
Çağrı Çöltekin, University of Tübingen

IWPT Chairs

Kenji Sagae, University of California, Davis
Stephan Oepen, University of Oslo

QUASY Chairs

Xinying Chen, University of Ostrava
Yaqin Wang, Guangdong University of Foreign Studies

Publication Chair

Sarah Jablotschkin, University of Hamburg

Local SyntaxFest 2025 Organizing Committee

Kaja Dobrovoljc, University of Ljubljana, SDJT
Špela Arhar Holdt, University of Ljubljana
Luka Terčon, University of Ljubljana
Marko Robnik-Šikonja, University of Ljubljana
Matej Klemen, University of Ljubljana
Sara Kos, University of Ljubljana
Timotej Knez, University of Ljubljana, SDJT
Tinca Lukan, University of Ljubljana

Special Thanks for designing the SyntaxFest 2025 logo to

Kim Gerdes, Université Paris-Saclay

Program Committee

Shared Program Committee

V.S.D.S.Mahesh Akavarapu, Eberhard-Karls-Universität Tübingen
Leonel Figueiredo de Alencar, Federal University of Ceará (UFC)
Patricia Amaral, Indiana University
Giuseppe Attardi, University of Pisa
John Bauer, Stanford University
David Beck, University of Alberta
Laura Becker, Albert-Ludwigs-Universität Freiburg
Aleksandrs Berdicevskis, Gothenburg University
Ann Bies, University of Pennsylvania
Igor Boguslavsky, Universidad Politécnica de Madrid
Bernd Bohnet, Google
Cristina Bosco, University of Turin
Gosse Bouma, University of Groningen
Miriam Butt, Universität Konstanz
G. A. Celano, Universität Leipzig
Heng Chen, Guangdong University of Foreign Studies
Xinying Chen, University of Ostrava
Jinho D. Choi, Emory University
Çağrı Çöltekin, University of Tuebingen
Daniel Dakota, Leidos
Stefania Degaetano-Ortlieb, Universität des Saarlandes
Kaja Dobrovoljc, University of Ljubljana
Jakub Dotlacil, Utrecht University
Gülşen Eryigit, Istanbul Technical University
Kilian Evang, Heinrich Heine University Düsseldorf
Pegah Faghiri, CNRS
Ramon Ferrer-i-Cancho, Universidad Politécnica de Catalunya
Marcos Garcia, Universidade de Santiago de Compostela
Kim Gerdes, Université Paris-Saclay
Loïc Grobol, Université Paris Nanterre
Bruno Guillaume, INRIA
Carlos Gómez-Rodríguez, Universidade da Coruña
Eva Hajicova, Charles University
Dag Trygve Truslew Haug, University of Oslo
Santiago Herrera, University of Paris Nanterre
Richard Hudson, University College London
Maarten Janssen, Charles University Prague
Jingyang Jiang, Zhejiang University
Mayank Jobanputra, Universität des Saarlandes
Sylvain Kahane, Université Paris Nanterre
Václava Kettnerová, Charles University Prague
Sandra Kübler, Indiana University
Guy Lapalme, University of Montreal
François Lareau, Université de Montréal
Miryam de Lhoneux, KU Leuven
Zoey Liu, University of Florida

Teresa Lynn, Dublin City University
 Jan Macutek, Slovak Academy of Sciences
 Robert Malouf, San Diego State University
 Marie-Catherine de Marneffe, UCLouvain
 Nicolas Mazziotta, Université de Liège
 Alexander Mehler, Johann Wolfgang Goethe Universität Frankfurt am Main
 Maitrey Mehta, University of Utah
 Wolfgang Menzel, Universität Hamburg
 Marie Mikulová, Charles University
 Aleksandra Miletić, University of Helsinki
 Jasmina Milićević, Dalhousie University
 Simon Mille, Dublin City University
 Yusuke Miyao, The University of Tokyo
 Noor Abo Mokh, Indiana University
 Simonetta Montemagni, Institute for Computational Linguistics “A. Zampolli” (ILC-CNR)
 Jiří Mírovský, Charles University Prague
 Kaili Müürisep, Institute of computer science, University of Tartu
 Anna Nedoluzhko, Charles University Prague
 Ruochen Niu, Beijing Language and Culture University
 Joakim Nivre, Uppsala University
 Stephan Oepen, University of Oslo
 Timothy John Osborne, Zhejiang University
 Petya Osenova, Sofia University “St. Kliment Ohridski”
 Agnieszka Patejuk, Polish Academy of Sciences
 Lucie Poláková, Charles University Prague
 Prokopis Prokopidis, Athena Research Center
 Mathilde Regnault, Universität Stuttgart
 Kateřina Rysová, University of South Bohemia
 Magdaléna Rysová, Charles University Prague
 Tanja Samardžić, University of Zurich
 Giuseppe Samo, Beijing Language and Culture University
 Haruko Sanada, Rissho University
 Nathan Schneider, Georgetown University
 Djamé Seddah, Sorbonne University
 Anastasia Shimorina, Orange
 Maria Simi, University of Pisa
 Achim Stein, University of Stuttgart
 Daniel G. Swanson, Indiana University
 Luka Terčon, Faculty of Arts, University of Ljubljana
 Giulia Venturi, Institute for Computational Linguistics “A. Zampolli” (ILC-CNR)
 Veronika Vincze, University of Szeged
 Yaqin Wang, Guangdong University of Foreign Studies
 Pan Xiaxing, Huaqiao University
 Chunshan Xu, Anhui Jianzhu University
 Nianwen Xue, Brandeis University
 Jianwei Yan, Zhejiang University
 Zdeněk Zabokrtský, Faculty of Mathematics and Physics, Charles University Prague
 Eva Zehentner, University of Zurich
 Amir Zeldes, Georgetown University
 Daniel Zeman, Charles University Prague
 Šárka Zikánová, Charles University Prague

Keynote

Subject prominence revisited: What makes entities salient?

Amir Zeldes
Georgetown University



Abstract: In this talk, I'll explore what makes certain entities stand out in discourse — what we might call more or less “salient” — and how speakers systematically identify them. Building on existing approaches to information structural “aboutness”, subjecthood, Centering Theory and animacy hierarchies, I argue that salience goes beyond surface categories such as definiteness, pronominalization and grammatical function. It's also shaped by deeper structures: distributional cues, discourse relations, hierarchical organization, genre conventions, and the communicative goals we infer from context. To get at this, I use a graded notion of salience based on how often entities are included in multiple human-written summaries of a text or conversation. Drawing on manually treebanked data from 24 different spoken and written genres in English, I ask: how is salience expressed for each entity mentioned in a discourse? I'll show that while traditional linguistic markers of salience all correlate with our salience scores to some extent, every rule has exceptions, and no single feature tells the whole story. Instead, salience cuts across all levels of linguistic structure, and the most informative theoretical model of the phenomenon must therefore combine cues from across morphosyntax, discourse structure, and functional pragmatics.

Bio: Amir Zeldes is Associate Professor of Computational Linguistics at Georgetown University, where he runs the Georgetown University Corpus Linguistics lab, Corpling@GU. He has worked on multilayer treebank construction and evaluation, including development of the Georgetown University Multilayer corpus (GUM) and datasets for low resource languages, such as the UD Coptic Treebank. His main area of research is computational discourse modeling, working on frameworks such as Enhanced Rhetorical Structure Theory (eRST) and Graded Salience, as well as topics such as coreference resolution, genre variation and summarization. He is currently president of the ACL Special Interest Group on Annotation (SIGANN).

Non-Archival Abstract

Segmentation of Sino-origin words to enhance the representation of Korean and Japanese in S/UD-format treebanks

Raoul Blin¹ and Jinnam Choi²

¹CNRS-CRLAO

²CLLE, Université Jean-Jaurès

In the Japanese and Korean S/UD treebanks, Chinese-origin words composed of two morphophonological units are not segmented, even when they are semantically transparent. We propose segmenting and annotating these words with dependency relations in order to achieve a more fine-grained and unified description of both languages. As an example, we apply this analysis to the pre-annotated GSD corpora in SUD format, and we examine the benefits and limitations of a rule-based approach.

Table of Contents

<i>Annotation of Chinese Light Verb Constructions within UMR</i>	
Jingyi Li, Jin Zhao, Nianwen Xue and Shili Ge	1
<i>Universal Dependencies for the Alemannic Alsatian Dialects</i>	
Barbara Hoff, Nathanaël Beiner and Delphine Bernhard	10
<i>Expanding the Universal Dependencies Ancient Hebrew Treebank with Constituency Data</i>	
Daniel G. Swanson	23
<i>Graph Databases for Fast Queries in UD Treebanks</i>	
Niklas Deworetzki and Peter Ljunglöf	32
<i>STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis</i>	
Luka Krsnik and Kaja Dobrovoljc	44
<i>«Are you Afraid of Ghosts?» A Proposal for Busting Predicate Ellipsis in Universal Dependencies</i>	
Claudia Corbetta, Federica Iurescia and Marco Carlo Passarotti	52
<i>Case Syncretism in Kasavakan Puyuma: A Field Data Analysis of Noun Phrase Markers</i>	
Deborah Watty, Yung-Jui Yao and Jens N. Watty	64
<i>Automatic Evaluation of Linguistic Validity in Japanese CCG Treebanks</i>	
Asa Tomita, Hitomi Yanaka and Daisuke Bekki	74
<i>Metaphorical Heads and Literal Dependents: Syntactic Properties of Metaphors in German</i>	
Stefanie Dipper	81
<i>A New Hebrew Universal Dependency Treebank: The First Treebank of Post-Rabbinic Historical Hebrew</i>	
Rachel Tal, Shlomit Fuchs, Orly Albeck, Elisheva Brauner, Yitzchak Lindenbaum, Ephraim Meiri and Avi Shmidman	91
<i>Syntax of referents of relative markers: Evidence from a corpus of learner English</i>	
Izabela Czerniak and Debopam Das	97
<i>An intonosyntactic treebank for spoken French: What is new with Rhapsodie?</i>	
Maria Paz Botero-Garcia, Emmett Strickland, Bruno Guillaume, Sylvain Kahane and Anne Lacheret-Dujour	111
<i>How to Create Treebanks without Human Annotators – An Indigenous Language Grammar Checker for Treebank Construction</i>	
Linda Wiecheteck, Flammie A Pirinen and Maja Lisa Kappfjell	119
<i>ComparaTree: A Multi-Level Comparative Treebank Analysis Tool</i>	
Luka Terčon and Kaja Dobrovoljc	129
<i>Universal Dependency Treebank for a low-resource Dardic Language: Torwali</i>	
Naeem Uddin and Daniel Zeman	140
<i>Legal-CGEL: Analyzing Legal Text in the CGELBank Framework</i>	
Brandon Waldon, Micaela Wells, Devika Tiwari, Meru Gopalan and Nathan Schneider	148
<i>Status of morphosyntactic features Illustration with written and spoken French UD treebanks</i>	
Sylvain Kahane, Bruno Guillaume, Léna Brun and Simeng Song	154