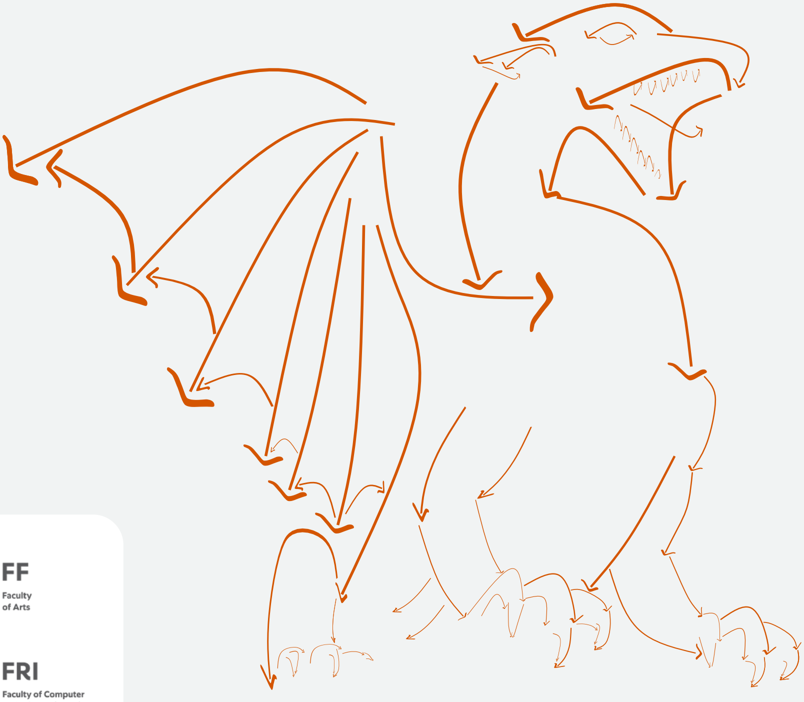
SYNTAXFEST 2025

5 Events for 1 Fest of Empirical Syntax

Ljubljana, 26 to 29 August 2025

Conference Guide



UNIVERSITY OF LJUBLJANA | FF Faculty of Arts

UNIVERSITY OF LJUBLJANA | FRI Faculty of Computer and Information Science

SDJT Slovensko društvo za jezikovne tehnologije

SyntaxFest 2025
5 Events for 1 Fest of Empirical Syntax

Ljubljana, 26 to 29 August 2025

Edited by: Kaja Dobrovoljc and Luka Terčon
Layout: Jure Preglau
Front page logo by: Kim Gerdes

Published by: Založba Univerze v Ljubljani (University of Ljubljana Press)
Issued by: Znanstvena založba Filozofske fakultete Univerze v Ljubljani
(University of Ljubljana Press, Faculty of Arts)
For the publisher: Gregor Majdič, Rector of the University of Ljubljana
For the issuer: Mojca Schlamberger Brezar, Dean of the Faculty of Arts, University of Ljubljana

Ljubljana, 2025
First e-edition. Digital copy of the book is available at: https://ebooks.uni-lj.si/zalozbaul/
DOI: 10.4312/9789612976361

Katаložni zapis o publikaciji (CIP) pripravili
v Narodni in univerzitetni knjižnici v Ljubljani
COBISS.SI-ID 245972739
ISBN 978-961-297-636-1 (PDF)

SyntaxFest 2025

5 Events for 1 Fest of Empirical Syntax

Ljubljana, 26 to 29 August 2025

Conference Guide

Venue:

Faculty of Law, University of Ljubljana
(Poljanski nasip 2, Ljubljana)

CONTENTS

INTRODUCTION

We are delighted to welcome you to SyntaxFest 2025 in Ljubljana, Slovenia. Continuing the tradition of previous editions in Paris (2019), Sofia (2021), and Washington DC (2023), SyntaxFest 2025 unites five independent yet closely connected events under one roof: the 18th International Conference on Parsing Technologies (IWPT 2025), the 8th Universal Dependencies Workshop (UDW 2025), the 8th International Conference on Dependency Linguistics (DepLing 2025), the 23rd Workshop on Treebanks and Linguistic Theories (TLT 2025), and the 3rd Workshop on Quantitative Syntax (QUASY 2025). Two pre-conference workshops organised by the COST Action CA21167 *Universality, Diversity and Idiosyncrasy in Language Technology* (UniDive) are also held in conjunction with the main event.

These events share a common focus on using corpora and treebanks to study syntax from both theoretical and computational perspectives, with growing emphasis on multilingual and cross-linguistic contexts. In addition to Slovenia being a fitting example of a small language community that has benefited from the openness, collaboration, and multilingual outlook fostered by this research community, hosting SyntaxFest in Ljubljana feels especially meaningful as the city was once home to the French linguist Lucien Tesnière (1893–1954), whose pioneering work laid the foundations of dependency grammar.

Against this backdrop, we extend our thanks to all who have contributed to making SyntaxFest 2025 possible. We are deeply grateful to the authors for bringing new ideas, analyses, and resources to the table; to our exceptional keynote speakers for taking the time to share their expertise and vision; and to the reviewers for their thoughtful work in shaping a high-quality program. Most importantly, we thank the workshop chairs for joining forces in this unique collaborative format, which has made it a truly community-driven effort.

Finally, we also thank our organising institutions for their support, our sponsors for making the event possible, the ACL Anthology for ensuring the proceedings are openly accessible, and our fellow organising committee members for their dedication behind the scenes.

Kaja Dobrovoljc
Chair of the SyntaxFest 2025 Local Organising Committee

CONFERENCE SCHEDULE

All conference sessions will take place at the Faculty of Law, University of Ljubljana (Poljanski nasip 2, Ljubljana). Talks will be held in the Red Hall, while poster sessions and coffee breaks will be organized in the lobby in front of the Red Hall.

The following pages provide an overview of the daily program, including workshops, keynotes, and social events. **For the most up-to-date version, visit the online program:** syntaxfest.github.io/syntaxfest25/programme.html.

TUESDAY, 26 August 2025	
	Session 1
14:00 – 14:30	**Conference Opening** Chair: Kaja Dobrovoljc
14:30 – 15:20	**Keynote: Isabel Papadimitriou (Harvard University)** What Can We Learn from Language Models? Chair: Stephan Oepen
15:20 – 15:50	Coffee Break
	Session 2 – IWPT Chair: Miryam de Lhoneux
15:50 – 16:10	**Step-by-step Instructions and a Simple Tabular Output Format Improve the Dependency Parsing Accuracy of LLMs** Hiroshi Matsuda, Chunpeng Ma, Masayuki Asahara
16:10 – 16:30	**An Efficient Parser for Bounded-Order Product-Free Lambek Categorial Grammar via Term Graph** Jinman Zhao, Gerald Penn
16:30 – 16:50	**Crosslingual Dependency Parsing of Hawaiian and Cook Islands Māori using Universal Dependencies** Gabriel H. Gilbert, Rolando Coto-Solano, Sally Akevai Nicholas, Lauren Houchens, Sabrina Barton, Trinity Pryor
16:50 – 17:10	**CCG Revisited: A Multilingual Empirical Study of the Kuhlmann-Satta Algorithm** Paul He, Gerald Penn
17:10 – 17:30	**High-Accuracy Transition-Based Constituency Parsing** John Bauer, Christopher D Manning
	Group Photo (Staircase)
17:30 – 20:00	Welcome Reception

WEDNESDAY, 27 August 2025

	Session 3 – UDW Chair: Gosse Bouma
09:00 – 09:50	**Keynote: Miryam de Lhoneux (KU Leuven)** Typologically informed NLP evaluation
09:50 – 10:10	**TreEn: A Multilingual Treebank Project on Environmental Discourse** Adriana Silvina Pagano, Patricia Chiril, Elisa Chierchiello, Cristina Bosco
10:10 – 10:30	**Crossing Dialectal Boundaries: Building a Treebank for the Dialect of Lesbos through Knowledge Transfer from Standard Modern Greek** Stavros Bompolas, Stella Markantonatou, Angela Ralli, Antonios Anastasopoulos
10:30 – 11:00	Coffee Break
	Session 4 – UDW Chair: Bruno Guillaume
11:00 – 11:15	**Negation in Universal Dependencies** Jamie Yates Findlay, Dag Trygve Truslew Haug
11:15 – 11:30	**A UD Treebank for Bohairic Coptic** Amir Zeldes, Nina Speransky, Nicholas E. Wagner, Caroline T. Schroeder
11:30 – 11:45	**Annotation of Relative Forms in the Egyptian-UJaen Treebank** Roberto A. Diaz Hernandez, Daniel Zeman
11:45 – 12:00	**MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs** Jaap Jumelet, Leonie Weissweiler, Arianna Bisazza
12:00 – 12:15	**Universal Dependencies for Suansu** Jessica K. Ivani, Kira Tulchynska
12:15 – 12:30	**Building UD Cairo for Old English in the Classroom** Lauren Levine, Junghyun Min, Amir Zeldes
12:30 – 14:00	Lunch (Cafeteria)
	Session 5 – UDW Chair: Dag Haug
14:00-14:15	**ShUD: the First Shanghainese Universal Dependency Treebank** Qizhen Yang
14:20-14:35	**Parallel Universal Dependencies Treebanks for Turkic Languages** Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Cagri Coltekin
14:35-14:50	**Towards better annotation practices for symmetrical voice in Universal Dependencies** Andrew Thomas Dyer, Colleen Alena O'Brien

14:50 – 15:10	**Annotating Second Language in Universal Dependencies: a Review of Current Practices and Directions for Harmonized Guidelines** Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna, Elena Volodina
15:10 – 15:30	**Reference and Modification in Universal Dependencies** Joakim Nivre, William Croft
15:30 – 16:00	Coffee Break
	Session 6 – Joint Poster Session A Chair: Cagri Coltekin
16:00 – 16:30	**Lightning talks (2 min per poster)** Location: Red Hall
16:30 – 17:30	**Poster Session** Location: Lobby

- **UD Treebanks for Esperanto as a natural language**
  Masanori Oya
- **UD-English-CHILDES: A Collected Resource of Gold and Silver Universal Dependencies Trees for Child Language Interactions**
  Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, Nathan Schneider
- **Universal Dependencies for Sindhi**
  John Bauer, Sakeena Shah, Muhammad Shaheer, Mir Afzal Ahmed Talpur, Zubair Sanjrani, Sarwat Qureshi, Shafi M Pirzada, Christopher D Manning, Mutee U Rahman
- **Universal Dependencies Treebank for Khoekhoe (KDT)**
  Kira Tulchynska, Sylvanus Job, Alena Witzlack-Makarevich
- **Extending the Enhanced Universal Dependencies – addressing subjects in pro-drop languages**
  Magali Sanches Duran, Elvis A. de Souza, Maria das Graças Volpe Nunes, Adriana Silvina Pagano, Thiago A. S. Pardo
- **Developing a Universal Dependencies Treebank for Alaskan Gwich'in**
  Matthew Kirk Andrews, Cagri Coltekin
- **Quid verbumst? Applying a definition of word to Latin in Universal Dependencies**
  Flavio Massimiliano Cecchini
- **Introducing KIParla Forest: seeds for a UD annotation of interactional syntax**
  Ludovica Pannitto, Eleonora Zucchini, Silvia Ballarè, Cristina Bosco, Caterina Mauri, Manuela Sanguinetti
- **Word Order Variation in Spoken and Written Corpora: A Cross-Linguistic Study of SVO and Alternative Orders**
  Nives Hüll, Kaja Dobrovoljc
- **A morpheme-based treebank for Gbaya, an Ubanguian language of Central Africa**
  Roulon-Doko Paulette, Sylvain Kahane, Bruno Guillaume
- **UD Annotation of Experience Clauses in Tigrinya**
  Michael Gasser, Nazareth Amlesom Kifle

18:00	**Guided Tour of Ljubljana**

THURSDAY, 28 August 2025	
	Session 7 – DepLing Chair: Joakim Nivre
09:00 – 09:50	**Keynote: Dan Zeman (Charles University, Prague)** Auxiliaries across Languages and Frameworks
09:50 – 10:10	**A corpus-driven description of OV order in Archaic Chinese** Qishen WU, Santiago Herrera, Pierre Magistry, Sylvain Kahane
10:10 – 10:30	**Periphrastic Verb Forms in Universal Dependencies** Lenka Krippnerová, Daniel Zeman
10:30 – 11:00	Coffee Break
	Session 8 – DepLing Chair: Bruno Guillaume
11:00 – 11:15	**Tracing Syntactic Complexity: Exploring the Evolution of Average Dependency Length Across Three Centuries of Scientific English** Marie-Pauline Krielke, Diego Alves, Luigi Talamo
11:15 – 11:30	**Modeling Syntactic Dependencies in Southern Dutch Dialects** Loic De Langhe, Jasper Degraeuwe, Melissa Farasyn, Veronique Hoste
11:30 – 11:45	**Assessing the Agreement Competence of Large Language Models** Alba Táboas García, Leo Wanner
11:45 – 12:00	**Genre Variation in Dependency Types: A Two-Level Genre Analysis Using the Czech National Corpus** Xinying Chen, Miroslav Kubát
12:00 – 12:15	**Distance and Projectivity as Predictors of Sentence Acceptability in Free Word Order Languages** Kirill Chuprinko, Artem Novozhilov, Arthur Stepanov
12:15 – 12:30	**Head-initial and head-Final coordinate structures in two annotation schemes of dependency grammar** Timothy John Osborne, Chenchen Song
12:30 – 14:00	Lunch (Cafeteria)
	Session 9 – TLT Chair: Heike Zinsmeister
14:00 – 14:50	**Keynote: Amir Zeldes (Georgetown University)** Subject prominence revisited: What makes entities salient?
14:50 – 15:10	**Legal-CGEL: Analyzing Legal Text in the CGELBank Framework** Brandon Waldon, Micaela Wells, Devika Tiwari, Meru Gopalan, Nathan Schneider
15:10 – 15:30	**Status of morphosyntactic features Illustration with written and spoken French UD treebanks** Sylvain Kahane, Bruno Guillaume, Léna Brun, Simeng Song
15:30 – 16:00	Coffee Break

	Session 10: Joint Poster Session B Chair: Stefanie Dipper
16:00 – 16:30	**Lightning talks (2 min per poster)** Location: Red Hall
16:30 – 17:30	**Poster Session** Location: Lobby

- **Universal Dependencies for the Alemannic Alsatian Dialects**
  Barbara Hoff, Nathanaël Beiner, Delphine Bernhard
- **Expanding the Universal Dependencies Ancient Hebrew Treebank with Constituency Data**
  Daniel G. Swanson
- **Graph Databases for Fast Queries in UD Treebanks**
  Niklas Deworetzki, Peter Ljunglöf
- **Segmentation of Sino-origin words to enhance the representation of Korean and Japanese in S/UD-format treebanks**
  Raoul Blin, Jinnam Choi
- **A New Hebrew Universal Dependency Treebank: The First Treebank of Post-Rabbinic Historical Hebrew**
  Rachel Tal, Shlomit Fuchs, Orly Albeck, Elisheva Brauner, Yitzchak Lindenbaum, Ephraim Meiri, Avi Shmidman
- **Universal Dependency Treebank for a low-resource Dardic Language: Torwali**
  Naeem Uddin, Daniel Zeman
- **Syntax of referents of relative markers: Evidence from a corpus of learner English**
  Izabela Czerniak, Debopam Das
- **A Typology of Non-Projective Patterns in Unas's and Teti's Pyramid Texts**
  Roberto A. Diaz Hernandez
- **Dependency Analysis of Chinese Comparative Sentences**
  Zexin Liu
- **Dative alternations in less-researched syntactic patterns of standard Croatian**
  Matea Andrea Birtić, Siniša Runjaić, Robert Sviben
- **A Quantitative Study of Subject-Predicate-Object Word Class Composition in vernacular Chinese Based on Dependency Grammar**
  Bingli Liu, Yiyi Zhao
- **Syntactic units and their length distributions: A case study in Czech**
  Michaela Nogolová, Michaela Koščová, Jan Macutek, Radek Cech
- **Modeling the Law of Abbreviation in Classical, Modern, and ChatGPT-Generated Chinese: A Power-Law Analysis of Structural Economy**
  Jianwei Yan, Heng Chen

19:00	**Conference Dinner**

FRIDAY, 29 August 2025	
	Session 11 – TLT Chair: Amir Zeldes
09:00 – 09:20	**ComparaTree: A Multi-Level Comparative Treebank Analysis Tool** Luka Terčon, Kaja Dobrovoljc
09:20 – 09:40	**Metaphorical Heads and Literal Dependents: Syntactic Properties of Metaphors in German** Stefanie Dipper
09:40 – 10:00	**Automatic Evaluation of Linguistic Validity in Japanese CCG Treebanks** Asa Tomita, Hitomi Yanaka, Daisuke Bekki
10:00 – 10:15	**Annotation of Chinese Light Verb Constructions within UMR** Jingyi Li, Jin Zhao, Nianwen Xue, Shili Ge
10:15 – 10:30	**STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis** Luka Krsnik, Kaja Dobrovoljc
10:30 – 11:00	Coffee Break
	Session 12 – QUASY Chair: Xinying Chen
11:00 – 11:50	**Keynote: Xiaofei Lu (The Pennsylvania State University)** The rhetorical and pragmatic functions of syntactically complex structures in academic and second language writing
11:50 – 12:10	**On the Flatness, Non-linearity, and Branching Direction of Natural Language and Random Constituency Trees: Analyzing Structural Variation within and across Languages** Taiga Ishii, Yusuke Miyao
12:10 – 12:30	**Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages** Santiago Herrera, Ioana-Madalina Silai, Bruno Guillaume, Sylvain Kahane
12:30 – 14:00	Lunch (Cafeteria)
	Session 13 – QUASY Chair: Jianwei Yan
14:00 – 14:15	**A Quantitative Study of Syntactic Complexity across Genres: Dependency Distance in English and Chinese** Yaqin Wang
14:15 – 14:30	**Syntactic Complexity in L2 Reading: A Comparison of Adapted and Original Czech Texts** Žaneta Stiborská, Michaela Nogolová, Xinying Chen, Miroslav Kubát
14:30 – 14:45	**First Insights into the Syntax of Slovene Student Writing: A Statistical Analysis of Šolar 3.0 vs. Učbeniki 1.0** Tina Munda, Špela Arhar Holdt

14:45 – 15:00	**Subject-Verb Agreement Alternations in Spanish Pseudopartitive Constructions: A Corpus Study** Marina Cerebrinsky
15:00 – 15:15	**A Computational Method for Analyzing Syntactic Profiles: The Case of the ELEXIS-WSD Parallel Sense-Annotated Corpus** Jaka Čibej
15:15 – 15:30	**Syntactic Complexity and News Credibility in Czech Media** Miroslav Kubát, Xinying Chen, Michaela Nogolová, Michal Místecký
15:30 – 16:00	Coffee Break
	Session 14: Joint Poster Session C Chair: Miroslav Kubát
16:00 – 16:20	**Lightning talks (2 min per poster)** Location: Red Hall
16:20 – 17:20	**Poster Session** Location: Lobby

- **Degree centrality as a measure of robustness of dependency structures of the sentences in a large-scale learner corpus of English**
  Masanori Oya
- **Application of Existing Readability Methods to the Ukrainian Language: A Comprehensive Study**
  Serhii D Prykhodchenko, Oksana Yu. Prykhodchenko
- **The Interplay of Noun Phrase Complexity and Modification Type in Scientific Writing**
  Isabell Landwehr
- **Predictability Effects of Spanish-English Code-Switching: A Directionality and Part of Speech Analysis**
  Josh Higdon, Valeria Pagliai, Zoey Liu
- **Do Multilingual Transformers Encode Paninian Grammatical Relations? A Layer-wise Probing Study**
  Akshit Kumar, Dipti Sharma, Parameswari Krishnamurthy
- **«Are you Afraid of Ghosts?» A Proposal for Busting Predicate Ellipsis in Universal Dependencies**
  Claudia Corbetta, Federica Iurescia, Marco Carlo Passarotti
- **Case Syncretism in Kasavakan Puyuma: A Field Data Analysis of Noun Phrase Markers**
  Deborah Watty, Yung-Jui Yao, Jens N. Watty
- **How to Create Treebanks without Human Annotators -- An Indigenous Language Grammar Checker for Treebank Construction**
  Linda Wiechetek, Flammie A Pirinen, Maja Lisa Kappfjell
- **An intonosyntactic treebank for spoken French: What is new with Rhapsodie?**
  Maria Paz Botero-Garcia, Emmett Strickland, Bruno Guillaume, Sylvain Kahane, Anne Lacheret-Dujour

17:20 – 17:30	**Closing Session**
18:00 - 18:30	**Guided Tour of National and University Library**

IWPT 2025

18th International Conference on Parsing Technologies

Abstracts

Keynote

Isabel Papadimitriou (Harvard University)

What Can We Learn from Language Models?

This talk will examine how linguistic theory can benefit from the recent surprising successes of language models in modeling human language production. Language models provide linguists with an unprecedented empirical tool to expand and test our theoretical hypotheses about language. I will go over two main methodologies for taking advantage of language models as an empirical tool. Firstly, examining language model internals as functional theories for how linguistic information can be represented in ways that lead to linguistic capabilities. Secondly, using model training as an empirical testbed, examining what kinds of environments make statistical language learning possible or harder. Both methodologies showcase the importance of developing empirical paradigms that narrow the gap between computational methods and linguistic concerns in order to make language models able to help us expand theoretical horizons.

Hiroshi Matsuda, Chunpeng Ma, Masayuki Asahara

Step-by-step Instructions and a Simple Tabular Output Format Improve the Dependency Parsing Accuracy of LLMs

Recent advances in large language models (LLMs) have enabled impressive performance in various tasks. However, standard prompting often struggles to produce structurally valid and accurate outputs, especially in dependency parsing. We propose a novel step-by-step instruction strategy, where universal part-of-speech tagging precedes the prediction of syntactic heads and dependency labels, and a simplified CoNLL-U like output format, our method achieves state-of-the-art accuracy on Universal Dependencies datasets across 17 languages without hallucination or contamination. We further show that multilingual fine-tuning simultaneously improves cross-language generalization performance. Our results highlight the effectiveness of explicit reasoning steps in LLM-based parsing and offer a scalable, format-consistent alternative to bracket-based approaches.

Jinman Zhao, Gerald Penn

An Efficient Parser for Bounded-Order Product-Free Lambek Categorial Grammar via Term Graph

Lambek Categorial Grammar (LCG) parsing has been proved to be an NP-complete problem. However, in the bounded-order case, the complexity can be reduced to polynomial time. Fowler (2007) first introduced the term graph, a simple graphical representation for LCG parsing, but his algorithm for using it remained largely inscrutable. Pentus (2010) later proposed a polynomial algorithm for bounded-order LCG parsing based on cyclic linear logic, yet both approaches remain largely theoretical, with no open-source implementations available. In this work, we combine the term-graph representation with insights from cyclic linear logic to develop a novel parsing algorithm for bounded-order LCG. Furthermore, we release our parser as an open-source tool.

Gabriel H. Gilbert, Rolando Coto-Solano, Sally Akevai Nicholas, Lauren Houchens, Sabrina Barton, Trinity Pryor

Crosslingual Dependency Parsing of Hawaiian and Cook Islands Māori using Universal Dependencies

This paper presents the first Universal Dependency (UD) treebank for 'Ōlelo Hawai'i (Hawaiian). We discuss some of the difficulties in describing Hawaiian grammar using UD, and train models for automatic parsing. We also combined this data with UD parses from another Eastern Polynesian language, Cook Islands Māori, to train a crosslingual Polynesian parser using UDPipe2. The crosslingual parser produced a statistically significant improvement of 2.4% in the labeled attachment score (LAS) when parsing Hawaiian, and this improvement didn't produce a negative impact in the LAS of Cook Islands Māori. We will use this parser to accelerate the linguistic documentation of Hawaiian.

Paul He, Gerald Penn

CCG Revisited: A Multilingual Empirical Study of the Kuhlmann-Satta Algorithm

We revisit the polynomial-time CCG parsing algorithm introduced by Kuhlmann & Satta (2014), and provide a publicly available implementation of it. We evaluate its empirical performance against a naive CKY-style parser across the Parallel Meaning Bank (PMB) corpus. While the fast parser is slightly slower on average, relative to the size of the PMB, but the trend improves as a function of sentence length, and the PMB is large enough to witness an inversion. Our analysis quantifies this crossover and highlights the importance of derivational context decomposition in practical parsing scenarios.

John Bauer, Christopher D Manning

High-Accuracy Transition-Based Constituency Parsing

Constituency parsers have improved markedly in recent years, with the F1 accuracy on the venerable Penn Treebank reaching 96.47, half of the error rate of the first transformer model in 2017. However, while dependency parsing frequently uses transition-based parsers, it is unclear whether transition-based parsing can still provide state-of-the-art results for constituency parsing. Despite promising work by Liu and Zhang in 2017 using an in-order transition-based parser, recent work uses other methods, mainly CKY charts built over LLM encoders. Starting from previous work, we implement self-training and a dynamic oracle to make a language-agnostic transition-based constituency parser. We test on seven languages; using Electra embeddings as the input layer on Penn Treebank, with a self-training dataset built from Wikipedia, our parser achieves a new SOTA F1 of 96.61.

UDW 2025

8th Universal Dependencies Workshop

Abstracts

Keynote

Miryam de Lhoneux (KU Leuven)

Typologically informed NLP evaluation

NLP has a long history of focusing mainly on English. While increasing efforts are being made towards making language technology more multilingual, English remains the language on which NLP technology is developed first, and applied to other languages next, which inevitably leads to degraded performance compared to English. This talk is about reversing this trend and putting multilinguality at the core of NLP, rather than at the periphery. I describe how typology can inform NLP evaluation, using our recently proposed language sampling framework. A strong limitation of the approach is the state of multilingual datasets, which tend to lack coverage, be machine-translated or have questionable quality. UD is an exception, and I emphasize the role it can play in establishing best practices in multilingual NLP evaluation.

Adriana Silvina Pagano, Patricia Chiril, Elisa Chierchiello, Cristina Bosco

TreEn: A Multilingual Treebank Project on Environmental Discourse

The increasing complexity of environmental discourse is directly proportional to the growing complexity of environmental debates present today in all communication media. While linguistic and communication studies have been pursued on this discourse, the development of computational linguistic tools and resources dedicated to support its analysis and interpretation is still very incipient. For one, no morphosyntactic resources specific to the environmental domain can be found on major platforms and repositories. This paper introduces TreEn, a multilingual treebank project in progress which compiles texts on environmental discourse produced in different conversational and communication contexts. In particular, it reports on the parallel component of the project and discusses issues faced during sentence-level alignment between original and translated texts, annotation of texts following UD guidelines, and labeling entities drawing on an ontology of environmental-related topics. This novel resource is expected to support environmental discourse analysis by providing morphological and syntactical data to enable cross-language and cross-cultural comparison based on the semantics of the entities annotated in the treebank.

Stavros Bompolas, Stella Markantonatou, Angela Ralli, Antonios Anastasopoulos

Crossing Dialectal Boundaries: Building a Treebank for the Dialect of Lesbos through Knowledge Transfer from Standard Modern Greek

This paper presents the first treebank for the dialect of Lesbos, a low-resource living Northern variety of Modern Greek (MG), annotated according to the Universal Dependencies (UD) framework. So far, the only dialectal treebank available for Greek developed with cross-dialectal knowledge transfer is an East Cretan one, which belongs to the same Southern branch as Standard Modern Greek (SMG). Our study investigates the effectiveness of cross-dialectal knowledge transfer between dialectologically less similar varieties of the same language by leveraging knowledge from SMG to annotate the Northern dialect of Lesbos. We describe the annotation process, present the resulting treebank, inject additional linguistic knowledge to enhance the results, and evaluate the effectiveness of cross-dialectal knowledge transfer for active annotation. Our findings contribute to a better understanding of how dialectal variation within language families affects knowledge transfer in the UD framework, with implications for other low-resource varieties.

Jamie Yates Findlay, Dag Trygve Truslew Haug

Negation in Universal Dependencies

In this paper we study the representation of negation in UD treebanks. We show that the existing annotations are often inconsistent with the guidelines and that there are ill-motivated differences in annotation of constructions across and even within languages. Moreover, we argue that even if the annotation of the two negation-related features (Polarity=Neg and PronType=Neg) were consistent, these two features would be inadequate for straightforwardly expressing the semantics of negation because they relate to the word level only and hence to form rather than meaning. We therefore propose to add two features, Negated=+ and DoubleNegated=+, which directly encode when a predicate is semantically under negation, and thereby allow a straightforward semantic interpretation of a UD parse in terms of negation.

Amir Zeldes, Nina Speransky, Nicholas E. Wagner, Caroline T. Schroeder

A UD Treebank for Bohairic Coptic

Despite recent advances in digital resources for other Coptic dialects, especially Sahidic, Bohairic Coptic, the main Coptic dialect for pre-Mamluk, late Byzantine Egypt, and the contemporary language of the Coptic Church, remains critically under-resourced. This paper presents and evaluates the first syntactically annotated corpus of Bohairic Coptic, sampling data from a range of works, including Biblical text, saints' lives and Christian ascetic writing. We also explore some of the main differences we observe compared to the existing UD treebank of Sahidic Coptic, the classical dialect of the language, and conduct joint and cross-dialect parsing experiments, revealing the unique nature of Bohairic as a related, but distinct variety from the more often studied Sahidic.

Roberto A. Diaz Hernandez, Daniel Zeman

Annotation of Relative Forms in the Egyptian-UJaen Treebank

Relative forms are a distinctive morphosyntactic feature of Earlier Egyptian. They pose a challenge when annotating them according to the Universal Dependencies approach. They are adjective finite verb forms, and therefore they have both verb and adjective properties, but they can also be used as nouns. The aim of this paper is to discuss the morphosyntactic methodology applied to their annotation in the Egyptian-UJaen treebank.

Jaap Jumelet, Leonie Weissweiler, Arianna Bisazza

MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs

We introduce MultiBLiMP 1.0, a massively multilingual benchmark of linguistic minimal pairs, covering 101 languages, 6 linguistic phenomena and containing more than 125,000 minimal pairs. Our minimal pairs are created using a fully automated pipeline, leveraging the large-scale linguistic resources of Universal Dependencies and UniMorph. MultiBLiMP evaluates abilities of LLMs at an unprecedented multilingual scale, and highlights the shortcomings of the current state-of-the-art in modelling low-resource languages.

Jessica K. Ivani, Kira Tulchynska

Universal Dependencies for Suansu

This contribution presents the Naga-Suansu Universal Dependencies (UD) treebank, the first resource of this kind for Suansu, an endangered and underdocumented Tibeto-Burman language spoken in Northeast India. This treebank follows the UD annotation framework. We describe the corpus composition, data sources, and annotation process, outlining the general structure of the treebank. In addition, we highlight morphosyntactic challenges where Suansu grammar does not fit neatly into the UD annotation schema and propose adaptations to better capture its structural proper- ties. As the first Tibeto-Burman language included in the UD project, the Naga-Suansu treebank serves several purposes: it contributes to the documentation and preservation of endangered languages, enables the understanding of cross-linguistic variation, and supports future research efforts in refining UD annotation practices for South and Southeast Asian languages.

Lauren Levine, Junghyun Min, Amir Zeldes

Building UD Cairo for Old English in the Classroom

In this paper we present a sample treebank for Old English based on the UD Cairo sentences, collected and annotated as part of a classroom curriculum in Historical Linguistics. To collect the data, a sample of 20 sentences illustrating a range of syntactic constructions in the world's languages, we employ a combination of LLM prompting and searches in authentic Old English data. For annotation we assigned sentences to multiple students with limited prior exposure to UD, whose annotations we compare and adjudicate. Our results suggest that while current LLM outputs in Old English do not reflect authentic syntax, this can be mitigated by post-editing, and that although beginner annotators do not possess enough background to complete the task perfectly, taken together they can produce good results and learn from the experience. We also conduct preliminary parsing experiments using Modern English training data, and find that although performance on Old English is poor, parsing on annotated features (lemma, hyperlemma, gloss) leads to improved performance.

Arianna Masciolini, Aleksandrs Berdicevskis, Maria Irena Szawerna, Elena Volodina

Annotating Second Language in Universal Dependencies: a Review of Current Practices and Directions for Harmonized Guidelines

Universal Dependencies (UD) is gaining popularity as an annotation standard for second language (L2) material. Grammatical errors and other interlanguage phenomena, however, pose significant challenges that official guidelines only address in part. In this paper, we give an overview of current annotation practices and provide some suggestions for harmonizing guidelines for learner corpora.

Joakim Nivre, William Croft

Reference and Modification in Universal Dependencies

Is the framework of Universal Dependencies (UD) compatible with findings from linguistic typology? To address this question, we need to systematically review how UD represents linguistic constructions in the world's languages, and how it handles the range of morphosyntactic variation attested in linguistic typology. In this paper, we start this review by discussing reference and modification constructions. The review shows that, although UD can represent all major constructions in this area, there are a number of cases where UD categories do not align systematically with a typological classification of constructions, and where constructional similarity is therefore not transparent across languages. We also identify limitations in the representation of certain morphosyntactic strategies, notably indexation and linkers. To overcome these limitations, we propose a number of revisions that may be considered for future versions of UD.

Masanori Oya

UD Treebanks for Esperanto as a natural language

This paper describes the details of UD-based morphological and syntactic annotations on Esperanto texts to construct its small-scale UD treebank. Though it was created as an international auxiliary language, Esperanto has increasingly been studied as a natural language both in linguistics and in NLP. This paper introduces the detail of manual annotation of UD morphological and relational tags and describes how the frequencies of these tags differ across the treebanks and discusses the possibility of future research of Esperanto as a natural language.

Xiulin Yang, Zhuoxuan Ju, Lanni Bu, Zoey Liu, Nathan Schneider

UD-English-CHILDES: A Collected Resource of Gold and Silver Universal Dependencies Trees for Child Language Interactions

CHILDES is a widely used resource of transcribed child and child-directed speech. This paper introduces UD-English-CHILDES, the first officially released Universal Dependencies (UD) treebank. It is derived from previously dependency-annotated CHILDES data, which we harmonize to follow unified annotation principles. The gold-standard trees encompass utterances sampled from 11 children and their caregivers, totaling over 48K sentences (236K tokens). We validate these gold-standard annotations under the UD v2 framework and provide an additional 1M silver-standard sentences, offering a consistent resource for computational and linguistic research.

John Bauer, Sakeena Shah, Muhammad Shaheer, Mir Afzal Ahmed Talpur, Zubair Sanjrani, Sarwat Qureshi, Shafi M Pirzada, Christopher D Manning, Mutee U Rahman

Universal Dependencies for Sindhi

Sindhi is an Indo-Aryan language spoken primarily in Pakistan and India by about 40 million people. Despite this extensive use, it is a low resource language for NLP tasks, with few datasets or pretrained embeddings available. In this work, we explore linguistic challenges for annotating Sindhi in the UD paradigm, such as language-specific analysis of adpositions and verb forms. We use this analysis to present a newly annotated dependency treebank for Universal Dependencies, along with pretrained embeddings and an annotation pipeline specifically for Sindhi annotation.

Kira Tulchynska, Sylvanus Job, Alena Witzlack-Makarevich

Universal Dependencies Treebank for Khoekhoe (KDT)

This paper reports on the development of the first dependency treebank for Khoekhoe (KDT). Khoekhoe (Khoe-Kwadi, Namibia) is a low-resource language with few linguistic and computational resources available publicly. This treebank consists of 29k words across six texts taken from various registers. It includes a substantial portion of spoken conversational data. These sentences were annotated manually according to the Universal Dependencies framework. In this paper, apart from presenting the strategies that have been followed to create the treebank, we also discussed some challenging morphological features and syntactic constructions found in the corpus and outlined how we have handled them using the current Universal Dependencies specification.

Arofat Akhundjanova, Furkan Akkurt, Bermet Chontaeva, Soudabeh Eslami, Cagri Coltekin

Parallel Universal Dependencies Treebanks for Turkic Languages

We introduce the first fully aligned and manually annotated parallel Universal Dependencies (UD) treebanks for four Turkic languages: Azerbaijani, Kyrgyz, Turkish, and Uzbek. These resources currently consist of 148 strategically selected sentences that illustrate typologically significant morphosyntactic phenomena across these related yet distinct languages. These parallel treebanks enable systematic comparative studies of Turkic syntax and may be instrumental in cross-lingual NLP applications. All treebanks are available as part of UD v2.16.

Andrew Thomas Dyer, Colleen Alena O’Brien

Towards better annotation practices for symmetrical voice in Universal Dependencies

Austronesian languages exhibit features that are challenging for Universal Dependencies: most notably, the symmetric voice system, whereby agent, patient, and instrumental arguments (among others) can be the pivot of a transitive structure – complicating the usual assumption that subjects of transitive sentences are semantic agents, and objects semantic patients. To showcase our ideas of how to address the representation of such systems in Universal Dependencies, we introduce a small treebank of sentences from texts and elicitation sessions in Gorontalo, an Austronesian language of Sulawesi (Indonesia), which exhibits a Philippine-type voice system. We discuss the annotation guidelines for this language, and the extensions of the Universal Dependencies guidelines that are needed to accommodate this and other Austronesian languages.

Magali Sanches Duran, Elvis A. de Souza, Maria das Graças Volpe Nunes, Adriana Silvina Pagano, Thiago A. S. Pardo

Extending the Enhanced Universal Dependencies – addressing subjects in pro-drop languages

Enhanced Universal Dependencies (EUD) serve as a crucial link between syntax and semantics. Beyond basic syntactic dependencies, EUD provides valuable refined logical connections for downstream tasks such as semantic role labeling, coreference resolution, information extraction, and question answering. The original EUD framework defines six types of relationships, but this paper introduces an extension designed to address subject propagation in pro-drop languages. This "Extended EUD" proposal increases the number of relationships that may be annotated in sentences, improving linguistic representation. Additionally, we report our experiments on a corpus of Portuguese (a pro-drop language), which we make publicly available to the research community.

Matthew Kirk Andrews, Cagri Coltekin

Developing a Universal Dependencies Treebank for Alaskan Gwich'in

This paper presents a Universal Dependencies (UD) treebank of Gwich'in, a severely endangered Athabascan language. The treebank, developed using instructional materials and dictionaries, includes 313 annotated sentences. This paper discusses the methodology used to construct the treebank, the linguistic challenges faced, and the implications of annotating a polysynthetic, morphologically complex language within the Universal Dependencies framework. The treebank was released with UD version 2.15 and available at https://github.com/UniversalDependencies/UD_Gwichin-TueCL/.

Flavio Massimiliano Cecchini

Quid verbumst? Applying a definition of word to Latin in Universal Dependencies

Words, more specifically “syntactic words”, are at the centre of a dependency-based approach like Universal Dependencies. Nonetheless, its guidelines do not make explicit how such a word should be defined and identified, and so it happens that different treebanks use different standards to this end. To counter this vagueness, the community has been recently discussing a definition put forward in (Haspelmath, 2023) which is not fully uncontroversial. This contribution is a preliminary case study that tries its hand at concretely applying this definition (except for compounds) to Latin in order to gain more insights about its operability and groundedness. This is helped by the spread of Latin over many treebanks, the presence of good linguistic resources to analyse it, and a linguistic type which is probably not fully considered in (Haspelmath, 2023). On the side, this work shows once more the difficulties of turning theoretical definitions into working directives in the realm of linguistic annotation.

Qizhen Yang

ShUD: the First Shanghainese Universal Dependency Treebank

This paper introduces ShUD, the first Universal Dependencies (UD) treebank for Shanghainese, a Wu Chinese variant spoken by approximately 14 million people but severely under-resourced in NLP. The treebank is built through a scalable annotation pipeline that exploits grammatical parallels between Shanghainese and Mandarin. Our pipeline also provides a practical strategy for bootstrapping resources for other Chinese dialects. We documented syntactic phenomena unique to Shanghainese within the UD framework and fine-tuned a dependency parser using our annotated treebank, contributing a foundation to both NLP tool development and cross-linguistic syntactic research.

DepLing 2025

8th International Conference on Dependency Linguistics

Abstracts

Keynote

Daniel Zeman (Charles University, Prague)

Auxiliaries across Languages and Frameworks

In my talk, I will discuss the status of auxiliaries (i.e., auxiliary verbs as well as uninflected non-verbal particles with auxiliary function) in dependency treebanks. The focus will be on two frameworks, Universal Dependencies (UD) and the Prague family of treebanks, rooted in the Functional Generative Description. However, I will occasionally show examples from other treebanks and frameworks, encountered during the HamleDT harmonization effort.

Besides looking at various treatments of auxiliaries in different annotation schemes, I will also discuss the question of delimiting the set of auxiliaries in individual languages (or, more exactly, the set of words that receive the special treatment in the respective annotation schemes). Various grammatical tests may be available, but sometimes the auxiliaries are simply enumerated by traditional school grammar. Moreover, there is a scale of categories ranging from pure grammatical auxiliaries through modals and phase verbs to various semantically bleached verbs that take other verbs as complements, yet their contribution is lexical rather than grammatical and their syntactic behavior shows no anomalies. All these aspects complicate finding a unified definition that would be applicable in a multi-lingual dataset, such as HamleDT or UD.

In the last part of the talk, I will show some examples of contrastive cross-linguistic studies that would benefit from comparably defined auxiliaries. I will show how we encourage comparability in UD using a common database of auxiliaries, and I will argue that it has the potential to become a useful resource of its own.

Ludovica Pannitto, Eleonora Zucchini, Silvia Ballarè, Cristina Bosco, Caterina Mauri, Manuela Sanguinetti

Introducing KIParla Forest: seeds for a UD annotation of interactional syntax

The present project endeavors to enrich the linguistic resources available for Italian by introducing KIParla Forest, a treebank for the KIParla corpus - an existing and well-known resource for spoken Italian. This article contextualizes the project, describes the treebank creation process and design choices, and highlights future plans for next improvements.

Nives Hüll, Kaja Dobrovoljc

Word Order Variation in Spoken and Written Corpora: A Cross-Linguistic Study of SVO and Alternative Orders

This study investigates word order variation in spoken and written corpora across five Indo-European languages: English, French, Norwegian (Nynorsk), Slovenian, and Spanish. Using Universal Dependencies treebanks, we analyze the distribution of six canonical word orders (SVO, SOV, VSO, VOS, OSV, OVS). Our results reveal that spoken language consistently exhibits greater word order flexibility than written language. This increased flexibility manifests as a decrease in the dominant SVO pattern and a rise in alternative orders, though the extent of this variation differs across languages. Morphologically rich languages such as Slovenian and Spanish show the most pronounced shifts, while English remains syntactically rigid across modalities. These findings support the claim that modality significantly affects syntactic realizations and highlight the need for typological studies to account for spoken data.

Roulon-Doko Paulette, Sylvain Kahane, Bruno Guillaume

A morpheme-based treebank for Gbaya, an Ubanguian language of Central Africa

In this paper, we present the first treebank for Gbaya, a language from the under-resourced Niger-Congo family. The language has a rich system of tonal morphemes and virtually no affixes. The dependency analysis is based on a morpheme-based tokenisation and the treebank is also distributed in word-based Universal Dependencies version. Several constructions are discussed in the paper: genitive construction, clause coordination, sentence particles, adverbial and relative clauses, serial verb constructions, reported speech, topicalization, and focalization.

Michael Gasser, Nazareth Amlesom Kifle

UD Annotation of Experience Clauses in Tigrinya

We are developing a treebank for Tigrinya within the Universal Dependency (UD) framework. UD proposes a set of universal grammatical relations to capture dependency relations between words in any language. However, for some classes of verbs it is not a straightforward matter to know what grammatical relations the verbs are categorized for. In this paper we discuss the decisions we have had to make for the annotation of arguments of experience verbs in the Semitic language Tigrinya, which exhibit a number of unusual morphosyntactic properties. We describe a classification of experience verb roots in the language, based on the various ways in which the core experiencer and stimulus arguments are realized syntactically and morphologically and on which valence-changing operations the roots permit. We supplement our analysis with data from a morphological analysis of a Tigrinya corpus.

Qishen WU, Santiago Herrera, Pierre Magistry, Sylvain Kahane

A corpus-driven description of OV order in Archaic Chinese

This paper presents a quantitative study of Object-Verb (OV) order in Archaic Chinese based on a Universal Dependencies (UD) treebanks. Treating word order as a binary choice (OV vs VO), we train a sparse logistic-regression classifier that selects the most salient syntactic features needed for an accurate prediction to investigate the specific syntactic contexts allowing OV word order and to identify to what extent do these factors favour this order. The ranked features are understood as interpretable rules, and their coverage and precision as quantitative properties of each rule. The approach confirms earlier qualitative findings (e.g. pronoun object fronting and negation favour OV) and uncovers new contrasts in word order between different reflexive pronouns. It also identifies annotation errors that we corrected in the final analysis, illustrating how the quantitative models, combined with fine-grained corpus analysis, can improve treebank quality. Our study demonstrates that lightweight machine-learning techniques applied to an existing syntactic resource can reveal fine-grained patterns in historical word order and this can be reapplied to other languages.

Lenka Krippnerová, Daniel Zeman

Periphrastic Verb Forms in Universal Dependencies

We propose a generalization of the morphological annotation in Universal Dependencies (UD) to phrases spanning multiple words, possibly discontinuous. Our focus area is that of periphrastic tenses, voices and other forms, typically consisting of a non-finite content verb combined with one or more auxiliaries; however, the same approach can be applied to other morphosyntactic constructions. We present a software tool that can detect periphrastic verb forms, extract the relevant morphological features from member words and combine them into new, phrase-level annotation. The tool currently detects periphrastic verb forms in 15 Slavic languages that are represented in UD and it is easily adaptable to other constructions and languages. Both the tool and the processed Slavic data are freely available.

Marie-Pauline Krielke, Diego Alves, Luigi Talamo

Tracing Syntactic Complexity: Exploring the Evolution of Average Dependency Length Across Three Centuries of Scientific English

We present a diachronic analysis of syntactic change in a corpus covering over 300 years (1665–1996) of scientific English, annotated with Universal Dependencies (UD) and Dependency Length (DL). We trace the development of average Dependency Length (aDL) as a measure of syntactic complexity in scientific English between 1665 and 1996. We describe the construction of the corpus and report on the evaluation of the UD annotation. We find that aDL initially decreases toward the 19th century, but then increases significantly in the 20th century. We show that this highly aggregate measure of aDL masks the underlying mechanisms driving changes in syntactic complexity. A more fine-grained analysis of the dependency relations involved in these changes reveals that the increasing use of (multi-word) compounds is a dominant source of long, leftward-expanded noun phrases. This leads to an expansion of syntactic dependencies both within and beyond the noun phrase. The results offer a new perspective on syntactic complexity, shifting the focus from the sentence level to the phrasal level.

Loic De Langhe, Jasper Degraeuwe, Melissa Farasyn, Veronique Hoste

Modeling Syntactic Dependencies in Southern Dutch Dialects

Dependency parsing of non-normative language varieties remains a challenge for modern NLP. While contemporary parsers excel at standardized languages, dialectal variation -- especially in function words, conjunctives, and verb clustering -- introduces syntactic ambiguity that disrupts traditional parsing approaches. In this paper, we conduct a quantitative evaluation of syntactic dependencies in Southern Dutch dialects, leveraging a standardized dialect corpus to isolate syntactic effects from lexical variation. Using a neural biaffine dependency parser with various mono- and multilingual transformer-based encoders, we benchmark parsing performance on standard Dutch, dialectal data, and mixed training sets. Our results demonstrate that incorporating dialect-specific data significantly enhances parsing accuracy, yet certain syntactic structures remain difficult to resolve, even with dedicated adaptation. These findings highlight the need for more nuanced parsing strategies and improved syntactic modeling for non-normative language varieties.

Alba Táboas García, Leo Wanner

Assessing the Agreement Competence of Large Language Models

While the competence of LLMs to cope with agreement constraints has been widely tested in English, only a very limited number of works deals with morphologically rich(er) languages. In this work, we experiment with 25 mono- and multilingual LLMs, applying them to a collection of more than 5,000 test examples that cover the main agreement phenomena in three Romance languages (Italian, Portuguese, and Spanish) and one Slavic Language (Russian). We identify which of the agreement phenomena are most difficult for which models and challenge some common assumptions of what makes a good model. The test suites into which the test examples are organized are openly available and can be easily adapted to other agreement phenomena and other languages for further research.

Xinying Chen, Miroslav Kubát

Genre Variation in Dependency Types: A Two-Level Genre Analysis Using the Czech National Corpus

This paper examines how dependency type distributions vary across genres in the Czech National Corpus (SYN2020). Using a two-level genre classification, broad categories and fine-grained subgenres, we identify genre-sensitive syntactic patterns through relative frequency analysis. The results show that some dependency types (e.g. Atr 'attribute') vary consistently across genres, while others (e.g. ExD 'part of discourse ellipsis') show sensitivity only at the subgenre level. Our dependency-based approach extends common multidimensional analyses based on lexical-grammatical co-occurrences, directly capturing syntactic evidence and improving interpretability. Our findings also highlight the importance of fine-grained genre distinctions in revealing syntactic variation.

Kirill Chuprinko, Artem Novozhilov, Arthur Stepanov

Distance and Projectivity as Predictors of Sentence Acceptability in Free Word Order Languages

This study investigates how two core metrics rooted in Dependency Grammar, Minimal Dependency Distance (MDD) and projectivity, predict sentence acceptability in Russian and Serbo-Croatian. Using exhaustive word order permutations in controlled five-word sentences, we model how these metrics relate to acceptability judgments in two psycholinguistic experiments. While MDD has been widely studied as a processing constraint, projectivity violations have received less attention in acceptability modeling. We show that both significantly affect judgments, with projectivity playing a surprisingly strong role. In addition, Serbo-Croatian's rigid clitic placement provides a natural test case for disentangling grammatical from processing constraints. Our findings offer a computationally precise, dependency-based model of acceptability that advances cognitively grounded language modeling for free word order languages.

Timothy John Osborne, Chenchen Song

Head-initial and head-Final coordinate structures in two annotation schemes of dependency grammar

The Universal Dependencies (UD) and Surface-Syntactic Universal Dependencies (SUD) annotation schemes view coordinate structures as head-initial. This contribution argues that a more flexible approach to coordinate structures is linguistically motivated, one that sees coordinate structures as head-initial in greater head-initial structures and as head-final in greater head-final structures. Support for this flexible approach comes from two areas: dependency distance and a nearness effect. In addition, two arguments that have been produced supporting the strictly head-initial approach are examined and refuted.

Roberto A. Diaz Hernandez

A Typology of Non-Projective Patterns in Unas's and Teti's Pyramid Texts

Abstract: The aim of this paper is to study the use of non-projective structures in Unas's and Teti's Pyramid Texts (ca. 2321–2279 BC) annotated in the Egyptian-UJaen treebank. It offers the first typology of non-projective patterns in Old Egyptian, and it discusses the causes for non-projectivity in the Old Egyptian language of Unas's and Teti's Pyramid Texts to conclude that non-projectivity is an exceptional phenomenon in these texts.

Zexin Liu

Dependency Analysis of Chinese Comparative Sentences

This paper examines the dependency structures of comparative sentences across various Chinese dialects. In equality comparatives, the comparative result is post-posed (R-back) in all Chinese dialects, which contrasts with English. Although Mandarin also follows the R-back pattern for superiority comparatives, dialects such as Hong Kong Cantonese and Southern Min adopt an R-front type, similar to English. However, Southern Min lacks a comparative marker, while English's comparative marker than dominates the standard of comparison. In contrast, the comparative marker in Cantonese does not dominate the standard. Through the calculation of dependency distances and syntactic tests, we argue that when the comparative result is preposed, it dominates the standard of comparison. Conversely, when the comparative construction fol-lows an R-back type, the comparative marker dominates the comparative result. This analysis aligns closely with the annotation choices of the Surface-Syntactic Universal Dependencies (SUD), differing significantly from those of the Universal Dependencies (UD) model.

Matea Andrea Birtić, Siniša Runjaić, Robert Sviben

Dative alternations in less-researched syntactic patterns of standard Croatian

Dative alternation in double object constructions is a frequently researched syntactic phenomenon, having been investigated across world languages. Consequently, even relatively smaller and under-resourced languages like Croatian have seen influential studies on the topic. Recent syntactic and semantic analyses of verbs in standard Croatian have identified less-explored instances of dative alternation. This contribution aims to describe the alternation between dative case and prepositional phrase for the non-agentive and intransitive uses of the verb služiti ('to serve'), as well as the dative alternation for the agentive and transitive uses of the verb izbjeći ('to avoid').

Bingli Liu, Yiyi Zhao

A Quantitative Study of Subject-Predicate-Object Word Class Composition in vernacular Chinese Based on Dependency Grammar

The paper aims at studying the evolution of lexical composition of subject-verb-object sentences in vernacular Chinese. Five corpora are constructed for the Tang and Five Dynasties, Song Dynasty, Yuan and Ming Dynasties, Qing Dynasty, and the present contemporary era which lasts for more than 1,000 years. The syntactic structures of these sentences are labeled, counted, and analyzed based on the theoretical foundation of dependency grammar, with the aim of investigating the evolution of the lexical category composition of the subject-predicate-object in vernacular Chinese over time. The results show that the ratio of nouns and pronouns in each period occupies the majority of the total number of subject lexemes, and the lexical composition of predicates has been very stable since ancient times, with verbal predicates accounting for the vast majority of predicates. Compared with the subject lexical composition, objects are richer and the lexical composition changes more slowly.

TLT 2025

23rd Workshop on Treebanks and Linguistic Theories

Abstracts

Keynote

Amir Zeldes (Georgetown University)

Subject prominence revisited: What makes entities salient?

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.

Luka Krsnik, Kaja Dobrovoljc

STARK: A Toolkit for Dependency (Sub)Tree Extraction and Analysis

We present STARK, a lightweight and flexible Python toolkit for extracting and analyzing syntactic (sub)trees from dependency-parsed corpora. By systematically slicing each sentence into interpretable syntactic units based on configurable parameters, STARK enables bottom-up, data-driven exploration of syntactic patterns at multiple levels of abstraction—from fully lexicalized constructions to general structural templates. It supports any CoNLL-U-formatted corpus and is available as a command-line tool, Python library, and interactive online demo, ensuring seamless integration into both exploratory and large-scale corpus workflows. We illustrate its functionality through case studies in noun phrase analysis, multiword expression identification, and syntactic variation across corpora, demonstrating its utility for a wide range of corpus-driven syntactic investigations.

Sylvain Kahane, Bruno Guillaume, Léna Brun, Simeng Song

Status of morphosyntactic features Illustration with written and spoken French UD treebanks

Morphosyntactic features used in UD treebanks have different status. If most of them correspond to values of inflectional morphemes, some describe lexical subclasses or are just conventional names of polysemic morphemes. Syncretism is also a challenge, because exact values are only deductible from contextual information. We propose an attempt at clarification and an implementation in the treebanks of written and spoken French.

Barbara Hoff, Nathanaël Beiner, Delphine Bernhard

Universal Dependencies for the Alemannic Alsatian Dialects

We present the first corpus of Alsatian Alemannic dialects following Universal Dependencies (UD) guidelines, a project which already covers many of the world's languages. Standard languages are represented to a greater extent than non-standard varieties in UD, and our corpus contributes to closing the gap in the lack of resources for Alsatian dialects by presenting the first UD treebank for these dialects, which are spoken in Northeastern France. Our corpus is annotated both with part-of-speech tags and dependency information, as well as French glosses and German lemmas, containing in total 975 sentences and 19,286 tokens, spanning over various text genres. In this article, we present our data, details of the annotation process, as well as some specific syntactic phenomena which differentiate and situate Alsatian with regards to both Standard German and some other German non-standard varieties. The addition of this corpus to the UD project allows for a higher visibility of the Alemannic Alsatian dialects in linguistic research, and provides a valuable resource for research in many fields, including NLP, syntax and comparative Germanic linguistics.

Daniel G. Swanson

Expanding the Universal Dependencies Ancient Hebrew Treebank with Constituency Data

This paper presents an effort to expand the annotation pipeline for the Ancient Hebrew Universal Dependencies treebank to make use of additional data, resulting in the addition of over 4000 sentences and roughly 100K words to the released treebank. The resulting treebank contains 5500 sentences and 145K words and the incorporation of converted constituency data has resulted in an annotation process which requires manual intervention in only around 15-20\% of sentences, even in previously unseen genres.

Niklas Deworetzki, Peter Ljunglöf

Graph Databases for Fast Queries in UD Treebanks

We investigate if labeled property graphs, and graph databases, can be an useful and efficient way of encoding UD treebanks, to facilitate searching for complex syntactic phenomena. We give two alternative encodings of UD treebanks into the off-the-shelf graph database Neo4j, and show how to translate syntactic queries into the graph query language Cypher. Our evaluation shows that graph databases can improve query times by several orders of magnitude, compared to existing approaches.

Raoul Blin, Jinnam Choi

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

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.

Rachel Tal, Shlomit Fuchs, Orly Albeck, Elisheva Brauner, Yitzchak Lindenbaum, Ephraim Meiri, Avi Shmidman

A New Hebrew Universal Dependency Treebank: The First Treebank of Post-Rabbinic Historical Hebrew

The corpus of post-Rabbinic historical Hebrew is a foundational corpus of Jewish heritage, containing over a billion words of legal, hermeneutical, and philosophic texts (and more). However, because the linguistic norms of the corpus diverge so often from that of modern Hebrew, the corpus cannot be computationally analyzed with existing Hebrew parsers. In order to fill this lacuna, we present the first Universal Dependencies corpus of post-Rabbinic historical Hebrew. The corpus comprises over 11,800 words, and we are pleased to release it to the community.

Naeem Uddin, Daniel Zeman

Universal Dependency Treebank for a low-resource Dardic Language: Torwali

This paper presents and discusses the linguistic phenomena encountered in the development of the ongoing first ever universal dependency treebank for Torwali the Language. Torwali belongs to the Kohistani sub-group of Dardic Indo-Aryan languages, and is considered an endangered (Moseley, 2010) and indigenous language, which makes it extremely low-resourced in terms of linguistic and computational resources. With the aim of including Torwali in Universal Dependencies (UD) (de Marneffe et al. 2021), we are annotating a diverse set of example sentences for POS tags, features and dependency relations.

Izabela Czerniak, Debopam Das

Syntax of referents of relative markers: Evidence from a corpus of learner English

We investigate the referents of relative markers of English relative clauses, focusing on their syntactic role in the matrix clauses. The referents, unlike relative markers and related features, have compratively remained understudied. We examine the syntactic environments of the referents as part of a larger project, which develops the ICLE-RC, a corpus of learner English texts annotated for relative clauses and related phenomena (it-/pseudo-clefts, existential-relatives, etc.). The corpus derives from the International Corpus of Learner English (ICLE; Granger et al., 2020), and contains 144 academic essays, representing six L1 backgrounds – Finnish, Italian, Polish, Swedish, Turkish, and Urdu. We annotate those texts for over 900 relative clauses (and over 400 related phenomena), with respect to a wide array of lexical, syntactic, semantic, and discourse features. Results from our analysis show that the relativisation of referents varies according to their syntactic functions. The referents are also observed to interact with other RC-features, yielding systematic variations across different L1 backgrounds, (some of) which can potentially be attributed to the typological properties of the associated L1.

Luka Terčon, Kaja Dobrovoljc

ComparaTree: A Multi-Level Comparative Treebank Analysis Tool

ComparaTree is a tool for comparative treebank analysis that combines various methods of quantitative linguistic analysis to provide a general overview of the differences and similarities between two treebanks. The comparison tool covers a range of subfields of linguistic analysis, providing a summary of the differences and similarities in terms of the lexical diversity, n-gram diversity, part-of-speech and dependency relation proportions, syntactic complexity, and syntactic diversity. We explain the various quantitative analyses performed on every level along with the generation of graphical visualizations, which add value by enabling user-friendly comparisons at a glance. We exemplify the comparison process by presenting the results produced by the tool when comparing two treebanks from the Universal Dependencies collection.

Stefanie Dipper

Metaphorical Heads and Literal Dependents: Syntactic Properties of Metaphors in German

In this paper we examine the way metaphors are expressed in language. Our starting hypothesis is that the two expressions that are central to metaphor – namely the metaphorical expression and the expression that represents the target of the metaphorical transfer – typically stand in a syntactic dependency relation: metaphorical heads govern literal dependents. An analysis of German sermons with 30k words confirms that the hypothesis applies in 67% of the cases. 10% show the reverse relationship and in 23% there is a common ancestor.

Asa Tomita, Hitomi Yanaka, Daisuke Bekki

Automatic Evaluation of Linguistic Validity in Japanese CCG Treebanks

In natural language inference, the accuracy of systems based on compositional semantics depends on the quality of syntactic analysis, which in turn relies on linguistically valid training and evaluation data, typically provided by treebanks. However, conventional treebank evaluation metrics focus on data coverage and fail to assess the linguistic validity of syntactic structures. This paper proposes novel evaluation methods to enable automatic and multifaceted assessment of linguistic validity. We apply these methods to a Japanese treebank based on combinatory categorial grammar and report the evaluation results.

Jingyi Li, Jin Zhao, Nianwen Xue, Shili Ge

Annotation of Chinese Light Verb Constructions within UMR

This paper discusses the challenges of annotating predicate-argument structures in Chinese light verb constructions (LVCs) within the Uniform Meaning Representation (UMR) framework, a cross-linguistic extension of Abstract Meaning Representation (AMR). A central challenge lies in reliably identifying LVCs in Chinese and determining their appropriate representation in UMR. We analyze the linguistic properties of Chinese LVCs, outline annotation difficulties for these structures and related constructions, and illustrate these issues through concrete examples. Our analysis focuses specifically on LVC.full types, where the light verb serves solely to convey morphological features and aspectual information. We exclude LVC.cause types, in which the light verb introduces an additional argument (e.g., a causal agent or source) to the event or state denoted by the nominal predicate. To address the practical challenge of semantic role assignment in Chinese LVCs, we propose a dual-path annotation approach: due to the compositional nature of these constructions, we recommend independently annotating the argument structure of the nominal predicate while systematically encoding the grammatical attributes and relations introduced by the light verb.

Brandon Waldon, Micaela Wells, Devika Tiwari, Meru Gopalan, Nathan Schneider

Legal-CGEL: Analyzing Legal Text in the CGELBank Framework

We introduce Legal-CGEL, an ongoing treebanking project focused on syntactic analysis of legal English text in the CGELBank framework (Reynolds et al., 2022), with an initial focus on US statutory law. When it comes to treebanking for legal English, we argue that there are unique advantages to employing CGELBank, a formalism that extends a comprehensive—and authoritative—formal description of English syntax (the _Cambridge Grammar of the English Language_; Huddleston & Pullum, 2002). We discuss some analytical challenges that have arisen in extending CGELBank to the legal domain. We conclude with a summary of immediate and longer-term project goals.

Claudia Corbetta, Federica Iurescia, Marco Carlo Passarotti

«Are you Afraid of Ghosts?» A Proposal for Busting Predicate Ellipsis in Universal Dependencies

This paper addresses the representation of ellipsis in dependency syntax, proposing both a theoretical and a practical workflow for its analysis and annotation in treebanks, following the state-of-the-art Universal Dependencies framework. We discuss the challenges of annotating ellipsis, with a focus on predicate ellipsis and its representation in dependency treebanks, and emphasize the importance of accounting for such phenomena for syntactic analysis and machine learning applications. We present a case study based on the Italian-Old treebank, demonstrating the applicability of the proposed workflows and invite the community to participate in this initiative with their own languages.

Deborah Watty, Yung-Jui Yao, Jens N. Watty

Case Syncretism in Kasavakan Puyuma: A Field Data Analysis of Noun Phrase Markers

Previous research has reported differing patterns of case syncretism across three dialects of Puyuma, an Austronesian language of Taiwan (Nanwang, Katipul, Ulivelivek). This study presents a quantitative analysis of case syncretism of noun phrase markers and disambiguation strategies in the Kasavakan dialect. Our dataset comprises 377 sentences elicited from five speakers, which we annotated for voice, potential semantic ambiguity, word order, and case marking of different semantic roles. We find evidence for a high degree of syncretism between genitive and nominative markers, alongside a decline in the use of genitive forms, particularly for common definite nouns. Some overlap with oblique markers is also attested, suggesting varying degrees of case syncretism between speakers. Topicalization appears to be the most frequent disambiguation strategy, while the order of non-topicalized noun phrases does not seem to aid disambiguation. Other factors, including age and individual experiences may contribute to inter-participant variation. These findings contribute to a more complete understanding of case marking in Puyuma by adding new empirical data from the Kasavakan dialect, where patterns of syncretism and disambiguation differ from previously described varieties.

Linda Wiechetek, Flammie A Pirinen, Maja Lisa Kappfjell

How to Create Treebanks without Human Annotators -- An Indigenous Language Grammar Checker for Treebank Construction

Creating treebanks for low resource languages is an important task. However, low resource Indigenous language contexts have not only limited resources in terms of text data, but also limited human resources that are available for linguistic annotation. We suggest a work-around by applying a Constraint Grammar operated rule-based dependency parser to do the work of creating a marked-up treebank. However, due to a lot of noise, meaning spelling and grammatical errors in South Sámi written texts, this tool often fails to create complete and correct trees. As a fix to this, we created a grammar checking tool for the most common South Sámi grammatical error types, which improves the quality of the dependency parser significantly. As both literacy and normative standards for most Indigenous languages are much more recent than for majority languages, spelling and grammatical variation and errors are a common source of noise, and the application of a correction tool like ours can be useful in the construction of treebanks for these languages.

Maria Paz Botero-Garcia, Emmett Strickland, Bruno Guillaume, Sylvain Kahane, Anne Lacheret-Dujour

An intonosyntactic treebank for spoken French: What is new with Rhapsodie?

This paper presents a new format of the Rhapsodie Treebank, which contains both syntactic and prosodic annotations, offering a comprehensive dataset for the study of spoken French.This integrated format allow us for complex multilevel queries and open the way for the extraction of intonosyntactic studies.

QUASY 2025

3rd Workshop on Quantitative Syntax

Abstracts

Keynote

Xiaofei Lu (The Pennsylvania State University)

The rhetorical and pragmatic functions of syntactically complex structures in academic and second language writing

Previous studies of linguistic complexity in academic and second language (L2) writing has often focused on quantitative differences across different writer groups and/or longitudinal changes over time, without systematic attention to the rhetorical or pragmatic functions that complex forms are used to convey. This talk argues for the importance of and delineates the scope of the function dimension of linguistic complexity analysis in L2 writing research, reviews the methods and findings of emerging efforts on this dimension, and discusses how future L2 writing research could attend to this dimension.

Michaela Nogolová, Michaela Koščová, Jan Macutek, Radek Cech

Syntactic units and their length distributions: A case study in Czech

This study investigates the length distributions of syntactic units in Czech across multiple hierarchical levels: sentences, independent clauses, clauses, phrases, subphrases, and chunks. Using a diverse dataset – including Universal Dependency treebanks, presidential speeches, the Czech Bible, and random sample from corpora of modern Czech – the analysis examines whether lengths of these syntactic units follow consistent distributional patterns. Length is defined as the number of immediate subunits, and the distributions were modeled using the hyper-Poisson distribution. The results demonstrate that the hyper-Poisson model fits well distributions of length of all abovementioned syntactic units, pointing to a common principle underlying the organization of syntactic structure in Czech.

Jianwei Yan, Heng Chen

Modeling the Law of Abbreviation in Classical, Modern, and ChatGPT-Generated Chinese: A Power-Law Analysis of Structural Economy

This study investigates the Law of Abbreviation—the inverse relationship between word length and frequency—across Classical, Modern, and ChatGPT-generated Chinese. Using a tri-partite parallel corpus and a power-law model y = a*x^(-b), we analyze the relationship between word length and the average usage frequency of words within a given word length category to assess structural economy. Results confirm consistent Zipfian distribution across all text types, with high R2 values indicating strong model fit. However, the parameter b varies significantly: Classical Chinese shows the steepest decline, suggesting strong pressure for brevity; Modern Chinese exhibits a moderated pattern; ChatGPT-generated texts display the weakest pressure, prioritizing fluency over compression. These differences reflect evolving communicative priorities and reveal that while AI models can mimic statistical distributions, they underrepresent deeper structural pressures found in natural language evolution. This study offers new insights into lexical optimization and the parameter b offers a useful metric for comparing structural efficiency across modalities. Implications are discussed in relation to language modeling, cognitive economy, and the evolution of linguistic structure.

Taiga Ishii, Yusuke Miyao

On the Flatness, Non-linearity, and Branching Direction of Natural Language and Random Constituency Trees: Analyzing Structural Variation within and across Languages

Natural languages exhibit remarkable diversity in their syntactic structures. Previous research has investigated the cross-lingual differences in local structural features such as word order or dependency relations. However, considering structural variation within individual language, it remains unclear how such features influence the variation in the overall constituency tree structure and hence the structural variation across languages. To this end, we focus on the shape of constituency trees, analyzing the cross-lingual overlap in the distributions of flatness, non-linearity, and branching direction. While acknowledging that the findings may be influenced by the potential annotation idiosyncrasies across treebanks, the experiments quantitatively suggest that flatness and branching direction vary significantly across languages. As for non-linearity, the cross-lingual difference was relatively small, and the distributions tend to skew towards linear structures. Furthermore, comparison with randomly generated trees suggests that while phrase category and frequency information is crucial for reproducing the branching direction found in natural languages, non-linearity can be replicated reasonably well even without such information.

Santiago Herrera, Ioana-Madalina Silai, Bruno Guillaume, Sylvain Kahane

Extraction of Contrastive Rules from Syntactic Treebanks: A Case Study in Romance Languages

In this paper, we develop a data-driven contrastive framework to extract common and distinctive linguistic descriptions from syntactic treebanks. The extracted contrastive rules are defined by a statistically significant difference in precision and classified as common and distinctive rules across the set of treebanks. We illustrate our method by working on object word order using Universal Dependencies (UD) treebanks in 6 Romance languages: Brazilian Portuguese, Catalan, French, Italian, Romanian and Spanish. We discuss the limitations faced due to inconsistent annotation and the feasibility of conducting contrasting studies using the UD collection.

Yaqin Wang

A Quantitative Study of Syntactic Complexity across Genres: Dependency Distance in English and Chinese

This study investigates syntactic complexity in fiction and news genres by analyzing mean dependency distances (MDD) across controlled sentence lengths in English and Chinese corpora. Results show that English fiction exhibits greater MDD than news, while Chinese fiction shows the reverse. More complex syntactic structures, i.e., complex coordination structures, are found in English fiction texts than in news writing. In contrast, Chinese news writing relies more on nominal modification and prepositional phrases that create long-distance dependencies than fiction texts. These findings show deviations from uniform correlations between genre formality and syntactic complexity across languages.

Žaneta Stiborská, Michaela Nogolová, Xinying Chen, Miroslav Kubát

Syntactic Complexity in L2 Reading: A Comparison of Adapted and Original Czech Texts

This corpus-based study explores the syntactic complexity of adapted Czech texts designed for learners of Czech as a second language (L2). It investigates how syntactic complexity varies according to learner proficiency levels (A2, B1, B2) as defined by the Common European Framework of Reference for Languages (CEFR) and how these adapted texts differ from their original versions. Quantitative analyses using metrics such as average sentence length (ASL), average clause length (ACL), mean dependency distance (MDD), and mean hierarchical distance (MHD) demonstrate clear systematic simplifications in adapted texts at lower proficiency levels. At A2 and B1 levels, adapted texts were found to be significantly less syntactically complex compared to their original counterparts. However, these differences diminished notably at the B2 proficiency level, indicating a gradual alignment of adapted texts with native-level syntactic complexity as learner proficiency increased. These results underscore the importance of careful syntactic calibration in creating educational materials for language learners, highlighting implications for curriculum design, instructional methodologies, and materials development. The findings offer valuable insights for language educators and textbook authors aiming to optimize reading materials to support language acquisition effectively.

Tina Munda, Špela Arhar Holdt

First Insights into the Syntax of Slovene Student Writing: A Statistical Analysis of Šolar 3.0 vs. Učbeniki 1.0

This study investigates the syntactic features of Slovene student writing by comparing essays from the Šolar 3.0 corpus (ages 13–19; primary and secondary school levels) with textbook texts from the Učbeniki 1.0 corpus aligned to the same educational stages. We apply quantitative syntactic analysis at two complementary levels: clause-type frequency (coordination, parataxis, and four types of subordination) and tree-based syntactic complexity measures (number of clauses, clauses per T-unit, and maximum parse-tree depth). Results show that students heavily rely on coordination and specific subordinate clauses (especially object and adverbial), producing more clauses per sentence and per T-unit than textbooks. However, their sentences tend to exhibit flatter syntactic structures, with shallower embedding in primary school and only modest increases in tree depth by secondary school. These findings reveal a divergence between surface-level complexity and hierarchical depth, highlighting developmental trends and instructional targets in written syntactic maturity. We conclude by discussing implications for syntactic development and directions for future research.

Marina Cerebrinsky

Subject-Verb Agreement Alternations in Spanish Pseudopartitive Constructions: A Corpus Study

Pseudopartitive constructions, following the format N1-of-N2 (such as a group of students), are known to feature alternations in their subject-verb agreement patterns, either with the N1 or the N2. Through a corpus analysis, this study investigates the possibility of a correlation between the choice of N1/N2 as an agreement trigger and the semantic type of the N1, as well as the animacy status of the N2. Although a positive correlation was found for N1 semantic type, no statistically significant results emerged for N2 animacy.

Jaka Čibej

A Computational Method for Analyzing Syntactic Profiles: The Case of the ELEXIS-WSD Parallel Sense-Annotated Corpus

In the paper, we present an approach to comparing corpora annotated with dependency relations. The method relies on the compilation of syntactic profiles – numeric vectors representing the relative frequencies of different syntactic (sub)trees extracted automatically with the STARK 3.0 open-access dependency tree extraction tool. We perform the extraction on the ELEXIS-WSD Parallel Sense-Annotated Corpus, which has recently been published as version 1.2 with UD dependency relation annotations for 10 European languages. The corpus provides an additional resource for contrastive studies in quantitative syntax. In addition to presenting the corpus and conducting some proof-of-concept analyses, we discuss several other potential uses and improvements to the proposed approach.

Miroslav Kubát, Xinying Chen, Michaela Nogolová, Michal Místecký

Syntactic Complexity and News Credibility in Czech Media

This study examines how syntactic complexity varies across news articles differing in credibility, using a Czech-language corpus annotated with five credibility levels: credible, partially credible, misleading, manipulative, and unclassifiable. We apply a dependency parsing pipeline and compute five syntactic metrics measuring features such as sentence length, clause density, and hierarchical depth. Results show that manipulative texts are structurally the most complex, while misleading and unclassifiable texts are simpler and more fragmented. Credible texts display balanced complexity consistent with journalistic norms. These findings highlight the role of syntax in shaping rhetorical strategies and contribute to the linguistic understanding of news credibility.

Masanori Oya

Degree centrality as a measure of robustness of dependency structures of the sentences in a large-scale learner corpus of English

This paper examines the differences in the robustness of syntactic dependency structures in written English produced by learners of varying proficiency levels and by native English speakers. The robustness of these dependency structures is represented by their degree centralities, and corpus-based investigation revealed that learners with higher proficiency levels tend to produce sentences with lower degree centralities. This means that they produce more robust, and more embedded sentences. It is also revealed that the sentences produced by native speakers of English tend to produce more embedded sentences than non-native speakers.

Serhii D Prykhodchenko, Oksana Yu. Prykhodchenko

Application of Existing Readability Methods to the Ukrainian Language: A Comprehensive Study

The Ukrainian language currently lacks a well-developed framework for assessing text readability. This study addresses this gap by focusing on three key contributions. First, we present the creation of UkrTB, a Ukrainian-language corpus of texts categorized by reader age. Second, we conduct a statistical analysis of the corpus, evaluating key linguistic features such as sentence length, word complexity, and part-of-speech distribution. Third, we systematically assess the applicability of existing readability formulas, including Flesch, Flesch-Kincaid, Matskovskii, Pisarek, and Solnyshkina et al., to Ukrainian texts. Our findings indicate that readability models developed for English and other Slavic languages exhibit significant limitations when applied to Ukrainian. While some methods demonstrate partial correlation with expected readability levels, others produce inconsistent results, underscoring the need for a specialized readability metric tailored to Ukrainian. This work lays the foundation for further research in Ukrainian readability assessment and the development of language-specific models.

Isabell Landwehr

The Interplay of Noun Phrase Complexity and Modification Type in Scientific Writing

We investigate the interplay of noun phrase (NP) complexity and modification type, namely the choice between pre- and postmodification, using a corpus-based approach. Our dataset is the Royal Society Corpus (RSC, Fischer et al. 2020), a diachronic corpus of English scientific writing. We find that the number of dependents, length of the head noun and distance to the head noun's own syntactic head (typically the main verb) affect the likelihood of pre- vs. postmodification: NPs with more dependents are more likely to be premodified, NPs with a longer head noun and a head noun closer to its own head are more likely to be postmodified. In addition, we find an effect of syntactic role and definiteness as well as time: The likelihood of premodification over postmodification increases with time and subject NPs as well as indefinite NPs are more likely to be premodified than NPs in other syntactic roles or definite NPs.

Josh Higdon, Valeria Pagliai, Zoey Liu

Predictability Effects of Spanish-English Code-Switching: A Directionality and Part of Speech Analysis

Research on code-switching (CS), the spontaneous alternation between two or more languages within a discourse, remains relatively new and often limited by the use of elicited production tasks, with some exceptions leveraging naturalistic corpora. This study analyses the effects of language directionality and part-of-speech (POS) tags on Spanish-English CS production between corpus modalities and speech communities. We use data from two spoken corpora: Miami Bangor Corpus (MBC; N = 261,711) and Spanish in Texas Corpus (STC; N = 416,784), as well as the written LinCE Corpus (N=278,093). Bootstrap analyses indicate that Spanish serves as the matrix language (i.e., the most used) for MBC and LinCE, while English is for STC. Logistic regression analyses show that the particle-coordinating conjunction combination was the strongest POS predictor of a CS. The results suggest that corpus modality and the speech community affect matrix language proportions and that both previously attested and unseen POS combinations modulate the production of Spanish-English CS.

Akshit Kumar, Dipti Sharma, Parameswari Krishnamurthy

Do Multilingual Transformers Encode Paninian Grammatical Relations? A Layer-wise Probing Study

Large multilingual transformers such as XLM-RoBERTa achieve impressive performance on diverse NLP benchmarks, but understanding how they internally encode grammatical information remains challenging. This study investigates the encoding of syntactic and morphological information derived from the Paninian grammatical framework—specifically designed for morphologically rich Indian languages—across model layers. Using diagnostic probing, we analyze the hidden representations of frozen XLM-RoBERTa-base, mBERT, and IndicBERT models across seven Indian languages (Hindi, Kannada, Malayalam, Marathi, Telugu, Urdu, Bengali). Probes are trained to predict Paninian dependency relations (by edge probing) and essential morphosyntactic features (UPOS tags, Vibhakti markers). We find that syntactic structure (dependencies) is primarily encoded in the middle-to-upper-middle layers (layers 6–9), while lexical features peak slightly earlier. Although the general layer-wise trends are shared across models, significant variations in absolute probing performance reflect differences in model capacity, pre-training data, and language-specific characteristics. These findings shed light on how theory-specific grammatical information emerges implicitly within multilingual transformer representations trained largely on unstructured raw text.

ORGANIZATION

Organizing Institutions

The Faculty of Arts of the University of Ljubljana (UL FF) educates students and creates top-quality educators with open and critical thinking in the humanities and social sciences, as well as educating teachers in these fields. It pays special attention to strengthening the disciplines of national importance that co-create Slovenian identity.

The Faculty of Computer and Information Science of the University of Ljubljana (UL FRI) is Slovenia's leading educational and research institution for computer and information science. The Faculty's main function is educating undergraduate and graduate computer science experts of various profiles, as well as engaging in research work which generates new knowledge and uncovers solutions to contemporary problems.

The Slovenian Language Technologies Society (SDJT) was founded in 1998 and joins people working on language technologies from the scientific, educational or user perspectives. The activities of the SDJT are aimed at promoting the development of language technologies for the Slovenian language.

Local Organizing Committee

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

Programme Chairs

18th International Conference on Parsing Technologies (IWPT):

- Kenji Sagae (University of California, Davis)
- Stephan Oepen (University of Oslo)

8th Universal Dependencies Workshop (UDW):

- Gosse Bomma (University of Groningen)
- Çağrı Çöltekin (University of Tübingen)

8th International Conference on Dependency Linguistics (DepLing):

- Eva Hajičová (Charles University, Prague)
- Sylvain Kahane (Université Paris Nanterre)

23rd Workshop on Treebanks and Linguistic Theories (TLT):

- Heike Zinsmeister (University of Hamburg)
- Sarah Jablotschkin (University of Hamburg)
- Sandra Kübler (Indiana University)

3rd Workshop on Quantitative Syntax (QUASY):

- Xinying Chen (University of Ostrava)
- Yaqin Wang (Guangdong University of Foreign Studies)

Publication Chair

- Sarah Jablotschkin (University of Hamburg)

Programme Committee

- Mahesh Akavarapu (Eberhard-Karls-Universität Tübingen)
- Leonel Figueiredo de Alencar (Federal University of Ceará)
- 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)
- Giuseppe G. A. Celano (Universität Leipzig)
- Heng Chen (Guangdong University of Foreign Studies)
- Xinying Chen (University of Ostrava)
- Jinho D. Choi (Emory University)
- Cagri Coltekin (University of Tuebingen)
- Daniel Dakota (Leidos)
- Stefania Degaetano-Ortlieb (Universität des Saarlandes)
- Kaja Dobrovoljc (University of Ljubljana)
- Jakub Dotlacil (Utrecht University)
- Gülşen Eryiğit (Istanbul Technical University)

- Kilian Evang (Heinrich Heine University Düsseldorf)
- Pegah Faghiri (CNRS)
- Ramon Ferrer-i-Cancho (Universidad Politécnica de Cataluna)
- 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 Samardzic (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)
- Zdenek Zabokrtsky (Charles University Prague)
- Eva Zehentner (University of Zurich)
- Amir Zeldes (Georgetown University)
- Daniel Zeman (Charles University Prague)
- Šárka Zikánová (Charles University Prague)
- Heike Zinsmeister (Universität Hamburg)

Support

We gratefully acknowledge the support of the following institutions and organizations whose contributions have helped make SyntaxFest 2025 possible:

- Centre for Language Resources and Technologies at the University of Ljubljana (CJVT UL)
- Slovene Common Language Resources and Technology Infrastructure (CLARIN.SI)
- COST Action CA21167 - Universality, diversity and idiosyncrasy in language technology (UniDive)
- The Centre of Excellence in Artificial Intelligence for Digital Humanities (CoE AI4DH)
- City of Ljubljana
- Ljubljana Tourism
- General Representative of Flanders in Austria, Hungary, Czech Republic, Slovakia and Slovenia
- Vitasis d.o.o.
- Alpineon d.o.o.
- Amebis d.o.o.
- Ustanova patra Stanislava Škrabca

Acknowledgment

The organization of SyntaxFest 2025 was partially supported by the Slovenian Research and Innovation Agency through the research program Language Resources and Technologies for Slovene (P6-0411) and the projects SPOT: A Treebank-Driven Approach to the Study of Spoken Slovenian (Z6-4617) and LLM4DH: Large Language Models for Digital Humanities (GC-0002). Co-funded by the European Union HORIZON-WIDERA-2023-TALENTS-01-01 grant 101186647 - AI4DH.

https://syntaxfest.github.io/syntaxfest25/