



LREC 2026

**Workshop on Structured Linguistic Data
and Evaluation (SLiDE)**

Workshop Proceedings

Editors

**Erhard Hinrichs, Joakim Nivre, Petya Osenova
James Pustejovsky, Claus Zinn**

May 11, 2026

©ELRA Language Resources Association (ELRA), 2026

These proceedings are licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0)

These proceedings are available from the ACL Anthology at <https://aclanthology.org>.

ISBN 978-2-493814-73-9

Message from the Workshop Organizers

In the last ten years, significant advances in deep learning models and the development of Large Language Models (LLMs) have revolutionized the fields of computational linguistics (CL) and natural language processing (NLP). In turn, this has led to a complete re-assessment of the language resources and evaluation practices necessary for training LLMs and analyzing their outputs. In particular, the availability of very large amounts of unstructured data for training foundational models has come into focus, while the value of high-quality structured linguistic data with rich annotations at various levels of linguistic analysis has been downplayed by comparison. However, as CL and NLP practitioners engage further with LLMs and debate their strengths and weaknesses, the importance of high-quality, structured linguistic data has been re-emphasized.

The workshop can be seen as related to the Treebanks and Linguistic Theories (TLT) conference series and the more recent SyntaxFest venue. Over the years, these venues have provided a central forum for high-quality research on treebanks, syntactic theory, syntax-semantics interface, structured meaning representations, and annotated linguistic resources. With record participation in recent years, they demonstrate the vitality and relevance of this line of work. The Workshop on Structured Linguistic Data is conceived as both a continuation of this tradition and an adaptation to the new realities of an LLM-dominated research landscape. The workshop brings together researchers from these overlapping traditions to advance methods, resources, and practices for integrating structured linguistic data into the LLM era.

The papers collected in this volume reflect a vibrant and multifaceted engagement with these issues, demonstrating that structured representations are not a relic of the past, but an essential component of robust, interpretable, and theoretically grounded natural language processing. A recurring theme across the program is the need to rigorously capture the syntax-semantics interface and complex event dynamics. For instance, contributions to the workshop explore how abstract meaning representations can be extended to handle quantification, and how subevent structures serve as vital predictors for entity identity changes in procedural texts. Such approaches underscore the necessity of deep lexical semantic analysis for tracking states and understanding discourse well beyond surface-level pattern matching.

Equally prominent is the focus on the creation, curation, and analysis of specialized corpora and lexicons. The accepted papers highlight the immense linguistic diversity and complexity that foundational models still struggle to autonomously master. Researchers present novel findings on human label variation in coreference, the compilation of temporally faithful corpora for diachronic NLP, and domain-specific challenges, such as those found in specialized German sermons. Furthermore, there is a strong emphasis on word-internal structures and morphological phenomena, ranging from large-scale segmentation across dozens of languages to the intricacies of linking elements in German nominal compounding and non-concatenative morphology in Tashlhiyt Berber. By modeling these regularities, including the design of datasets for regular metaphor, the community continues to provide the high-fidelity data necessary to push computational linguistics forward.

The direct interplay between generative models and structured data forms another critical pillar of this year's proceedings. Several authors tackle the evaluation and enhancement of LLMs through structured lenses, whether by comparing natural and synthetic structured data regarding verb alternations, or by exploring corpus-grounded agentic LLMs for multilingual grammatical analysis. Additionally, methods for improving language models for lexicographic question answering and employing syntax as a semantic pivot in machine translation showcase practical pathways for injecting rich linguistic knowledge into contemporary pipelines.

Finally, methodological innovations in how we store, query, and utilize structured linguistic data continue to evolve. The program features exciting work on fast and flexible example-based treebank search using vector symbolic architectures, offering fresh mathematical and computational perspectives on representing and retrieving complex structures. Together with the insights from our keynote speakers, Shira Wein and Naiara Perez, the research presented in these proceedings charts a clear and exciting course. It is our hope that this volume not only documents the remarkable work being done at this critical intersection, but also inspires continued theoretical and practical development across the community.

The number of submissions was 30 which shows the relevance of the topic. From these, 21 papers were accepted – 7 for oral presentation and 14 as posters. The acceptance rate is thus 70 % and it indicates the high quality of the submissions.

The Organizers

Acknowledgements

We would like to thank our invited speakers Naiara Perez and Shira Wein for their stimulating keynotes and for their willingness to cover their own accommodations and travel expenses. Also, we are grateful to have had Michał Ulewicz (IBM) as secondary reviewer. We are also indebted for the guidance provided by the LREC workshop committee that made our preparations a smooth process.

Organizing Committee

Jan Hajič (Prague University, Czech Republic)
Erhard Hinrichs (Tübingen University, Germany)
Sandra Kübler (Indiana University, USA)
Joakim Nivre (Uppsala University, Sweden)
Petya Osenova (Sofia University and IICT-BAS, Bulgaria)
James Pustejovsky (Brandeis University, USA)

Program Committee

Meriem Beloucif (Uppsala University, Sweden)
Bernd Bohnet (Google, London, UK)
Johan Bos (University of Groningen, The Netherlands)
Gosse Bouma (University of Groningen, The Netherlands)
Çağrı Çöltekin (Tübingen University, Germany)
Daniel Dakota (Leidos, Indiana University, USA)
Miryam de Lhoneux (KU Leuven, Belgium)
Stephanie Dipper (Bochum University, Germany)
Katrin Erk (University of Massachusetts at Amherst, USA)
Kilian Evang (Düsseldorf University, Germany)
Kim Gerdes (Université Paris-Saclay, France)
Eva Hajičová (Charles University, Czech Republic)
Yoshihiko Hayashi (Waseda University, Japan)
Shu-Kai Hsieh (National Taiwan University, Taiwan)
Laura Kallmeyer (Düsseldorf University, Germany)
Ekaterina Kochmar (Mohamed Bin Zayed University of AI, Abu Dhabi)
Marie-Pauline Krielke (Saarland University, Germany)
Lori Levin (Carnegie Mellon University, USA)
Wolfgang Menzel (Hamburg University, Germany)
Marie Mikulová (Charles University, Czech Republic)
Kaili Müürisep (University of Tartu, Estonia)
John Nerbonne (Freiburg University, Germany)
Colleen Alena O'Brien (Palacky University, Olomouc, Czech Republic)
Stephan Oepen (Oslo University, Norway)
Lilja Øvrelid (University of Oslo, Norway)
Barbara Plank (LMU Munich, Germany)
Agata Savary (Université Paris-Saclay, France)
Kiril Simov (IICT, Bulgarian Academy of Sciences, Bulgaria)
Sanghoun Song (Korea University, Korea)
Sara Stymne (Uppsala University, Sweden)
Rob van der Goot (IT University of Copenhagen, Denmark)
Leonie Weissweiler (Uppsala University, Sweden)
Shuly Wintner (Haifa University, Israel)
Alina Wróblewska (Polish Academy of Sciences, Warsaw, Poland)
Nianwen Xue (Brandeis University, USA)
Daniel Zeman (Charles University, Czech Republic)
Heike Zinsmeister (Hamburg University, Germany)

Keynote Speakers

Naiara Perez (University of the Basque Country)
Shira Wein (Amherst College)

Table of Contents

<i>When Does Linguistics Help? Downstream Utility of Linguistic Frameworks in the Age of LLMs</i> Shira Wein	xv
<i>Building and Evaluating Instruction-Following LLMs for Low-Resource Languages</i> Naiara Perez	xvi
<i>Degrees of Subjectivity and Their Repercussions in Conversation. The View from Online Interactions</i> Gonzalo Freijedo Aduna, Anastasia Giannakidou and Alda Mari	1
<i>Constraints on Linking Element Choice in German Nominal Compounding: A Large-Scale Corpus Study</i> Maksim Shmalts	14
<i>Comparing Natural and Synthetic Structured Data: A Study of the Passive Verb Alternation in French and Italian</i> Giuseppe Samo and Paola Merlo	39
<i>Subevent Structure as a Predictor of Entity Identity Change in Procedural Text</i> Kyeongmin Rim and James Pustejovsky	52
<i>Fast and Flexible Example-Based Treebank Search with Vector Symbolic Architectures</i> Adam Roussel	64
<i>Modular Neural Machine Translation with a Semantic Pivot - Pilot Study Using AMR</i> Wenyang Gao, Yaxuan Li, Yunxin Bao, Shulin Huang and Yue Zhang	74
<i>Gutenberg+: A More Temporally Faithful Corpus for Diachronic NLP</i> Leon Hammerla and Alexander Mehler	86
<i>Automatic Lemmatisation for Norwegian</i> Ahmet Yildirim, Kristin Hagen and Dag Haug	93
<i>Quantification in Abstract Meaning Representation</i> Kiyong Lee and Chongwon Park	104
<i>Improving Slovene Language Models for Lexicographic Question Answering through Continued Pretraining and Instruction Fine-Tuning</i> Timotej Knez and Slavko Zitnik	114
<i>Structured Partial Predictability in Non-Concatenative Morphology: The Case of Tashlhiyt Berber</i> John Alderete and Hamza Sellami	124
<i>Towards Corpus-Grounded Agentic LLMs for Multilingual Grammatical Analysis</i> Matej Klemen, Tjaša Arčon, Luka Terčon, Marko Robnik-Sikonja and Kaja Dobrovoljc	136
<i>The L2 Network: A CEFR-Aligned Knowledge Graph for Grammar Domain Modeling</i> Luisa Ribeiro-Flucht and Xiaobin Chen	148

<i>Paraphrase Acquisition via Bilingual Pivoting Based on Neural Word Alignment</i> Risa Kondo, Seiji Sugiyama, Tomoyuki Kajiwara and Takashi Ninomiya	160
<i>Using syntax for the semantic representation of sentences</i> Iskandar Boucharenc, Eve Sauvage, Thomas Gerald, Julien Tourille, Sabrina Campano, Cyril Grouin and Sophie Rosset	169
<i>Modeling Word-Internal Structures: Morphological Segmentation Across 58 Languages</i> Vojtěch John, Zdeněk Žabokrtský and Benjamin Reeves	180
<i>DiNoS: Creating a Data-Driven German Noun Phrase Lexicon from Universal Dependencies</i> Jacob Lee Suchardt and Ronja Laarmann-Quante	191
<i>Figurative, Polysemous, Conventional: Designing a Dataset of Regular Metaphor</i> Anna Temerko, Pablo Gamallo and Marcos Garcia	202
<i>Domain-Specific Considerations in the Preparation of Specialized Corpora: A Case Study on a Corpus of German Sermons</i> Cora Haiber, Adam Roussel and Stefanie Dipper	212
<i>Semantic, Syntactic, Lexical: What Makes QA Augmentation Work in Limited Quantity?</i> Benedictus Kent Rachmat, Thomas Gerald, Takuya Nakamura, Zheng Zhang and Cyril Grouin	224
<i>Introducing corpora Hlava Cor and Hlava AD: Human Label Variation in Coreference and Dis- course Relations</i> Anna Nedoluzhko, Sarka Zikanova, Jiri Mirovsky, Milan Straka and Eva Hajicova	237

Workshop Program

Monday, May 11, 2026

- | | |
|--------------------|--|
| 09:00–09:10 | Opening
Room: 3
Chairs: Erhard Hinrichs, Joakim Nivre, Petya Osenova |
| 09:10–10:00 | Keynote 1: Shira B. Wein, Amherst College
Room: 3
Chair: James Pustejovsky |
| 10:00–10:30 | Session 1
Room: 3
Chair: James Pustejovsky |
| 10:00–10:30 | <i>Degrees of Subjectivity and Their Repercussions in Conversation. The View from Online Interactions</i>
Gonzalo Freijedo Aduna, Anastasia Giannakidou and Alda Mari |
| 10:30–11:00 | Coffee Break 1 |
| 11:00–13:00 | Session 2
Room: Room 3
Chair: Petya Osenova |
| 11:00–11:30 | <i>Constraints on Linking Element Choice in German Nominal Compounding: A Large-Scale Corpus Study</i>
Maksim Shmalts |
| 11:30–12:00 | <i>Comparing Natural and Synthetic Structured Data: A Study of the Passive Verb Alternation in French and Italian</i>
Giuseppe Samo and Paola Merlo |
| 12:00–12:30 | <i>Subevent Structure as a Predictor of Entity Identity Change in Procedural Text</i>
Kyeongmin Rim and James Pustejovsky |
| 12:30–13:00 | <i>Fast and Flexible Example-Based Treebank Search with Vector Symbolic Architectures</i>
Adam Roussel |

Monday, May 11, 2026 (continued)

13:00–14:00	Lunch Break
14:00–15:10	Poster Session Chair: Joakim Nivre
14:00–15:10	<i>Modular Neural Machine Translation with a Semantic Pivot - Pilot Study Using AMR</i> Wenyang Gao, Yaxuan Li, Yunxin Bao, Shulin Huang and Yue Zhang
14:00–15:10	<i>Gutenberg+: A More Temporally Faithful Corpus for Diachronic NLP</i> Leon Hammerla and Alexander Mehler
14:00–15:10	<i>Automatic Lemmatisation for Norwegian</i> Ahmet Yildirim, Kristin Hagen and Dag Haug
14:00–15:10	<i>Quantification in Abstract Meaning Representation</i> Kiyong Lee and Chongwon Park
14:00–15:10	<i>Improving Slovene Language Models for Lexicographic Question Answering through Continued Pretraining and Instruction Fine-Tuning</i> Timotej Knez and Slavko Zitnik
14:00–15:10	<i>Structured Partial Predictability in Non-Concatenative Morphology: The Case of Tashlhiyt Berber</i> John Alderete and Hamza Sellami
14:00–15:10	<i>Towards Corpus-Grounded Agentic LLMs for Multilingual Grammatical Analysis</i> Matej Klemen, Tjaša Arčon, Luka Terčon, Marko Robnik-Sikonja and Kaja Dobrovoljc
14:00–15:10	<i>The L2 Network: A CEFR-Aligned Knowledge Graph for Grammar Domain Modeling</i> Luisa Ribeiro-Flucht and Xiaobin Chen
14:00–15:10	<i>Paraphrase Acquisition via Bilingual Pivoting Based on Neural Word Alignment</i> Risa Kondo, Seiji Sugiyama, Tomoyuki Kajiwara and Takashi Ninomiya

Monday, May 11, 2026 (continued)

- 14:00–15:10 *Using syntax for the semantic representation of sentences*
Iskandar Boucharenc, Eve Sauvage, Thomas Gerald, Julien Tourille,
Sabrina Campano, Cyril Grouin and Sophie Rosset
- 14:00–15:10 *Modeling Word-Internal Structures: Morphological Segmentation
Across 58 Languages*
Vojtěch John, Zdeněk Žabokrtský and Benjamin Reeves
- 14:00–15:10 *DiNoS: Creating a Data-Driven German Noun Phrase Lexicon from Uni-
versal Dependencies*
Jacob Lee Suchardt and Ronja Laarmann-Quante
- 14:00–15:10 *Figurative, Polysemous, Conventional: Designing a Dataset of Regular
Metaphor*
Anna Temerko, Pablo Gamallo and Marcos Garcia
- 14:00–15:10 *Domain-Specific Considerations in the Preparation of Specialized Cor-
pora: A Case Study on a Corpus of German Sermons*
Cora Haiber, Adam Roussel and Stefanie Dipper
- 15:10–16:00 **Keynote 2: Naiara Perez, University of the Basque Country**
Room: 3
Chair: Petya Osenova
- 16:00–16:30 **Coffee Break 2**
- 16:30–17:30 **Session 3**
Room: 3
Chair: Jan Hajič
- 16:30–17:00 *Semantic, Syntactic, Lexical: What Makes QA Augmentation Work in
Limited Quantity?*
Benedictus Kent Rachmat, Thomas Gerald, Takuya Nakamura, Zheng
Zhang and Cyril Grouin
- 17:00–17:30 *Introducing corpora Hlava Cor and Hlava AD: Human Label Variation in
Coreference and Discourse Relations*
Anna Nedoluzhko, Sarka Zikanova, Jiri Mirovsky, Milan Straka and Eva
Hajicova

Monday, May 11, 2026 (continued)

17:30–18:00 Closing Discussion

Room: 3

Chairs: Jan Hajič, Erhard Hinrichs, Joakim Nivre, Petya Osenova,
James Pustejovsky