LANGUAGE, DATA and KNOWLEDGE 2025

Proceedings of the 5th Conference on **Language, Data and Knowledge**

UniorPress
Naples
2025

PROCEEDINGS OF THE 5TH CONFERENCE ON LANGUAGE, DATA AND KNOWLEDGE

EDITORS:

Mehwish Alam, Institut Polytechnique de Paris, France
Andon Tchechmedjiev, Institut Mines-Télécom | EuroMov Digital Health in Motion, France
Jorge Gracia, University of Zaragoza, Spain
Dagmar Gromann, University of Vienna, Austria
Maria Pia di Buono, University of Naples "L'Orientale", Italy
Johanna Monti, University of Naples "L'Orientale", Italy
Maxim Ionov, University of Zaragoza, Spain

ISBN 978-88-6719-333-2

UniorPress — University of Naples "L'Orientale"
Via Nuova Marina 59 — 80133 Napoli (Italy)

UniorPress

Foreword

This volume presents the proceedings of the 5th Conference on Language, Data and Knowledge held in Naples, Italy, from 9 to 11 September 2025. Language, Data and Knowledge (LDK) is a biennial conference series on matters of human language technology, data science, and knowledge representation, initiated in 2017 by a consortium of researchers from the Insight Centre for Data Analytics at the National University of Ireland, Galway (Ireland), the Institut für Angewandte Informatik (InfAI) at the University of Leipzig (Germany), and the Applied Computational Linguistics Lab (ACoLi) at Goethe University Frankfurt am Main (Germany). Since the beginning, it has received the continuous support of an international Scientific Advisory Committee of leading researchers in natural language processing, linked data and Semantic Web, language resources and digital humanities. This edition builds upon the success of the inaugural event held in Galway, Ireland, in 2017, the second LDK in Leipzig, Germany, in 2019, the third LDK in Zaragoza, Spain, in 2021 and the fourth edition in Vienna, Austria, in 2023. The LDK Conference was recognised and incorporated into the esteemed CORE ranking in 2022. This fifth edition of the LDK conference is hosted by the University of Naples "L'Orientale", Italy.

As a biennial event, LDK aims to bring together researchers from across disciplines concerned with acquiring, curating and using language data in the context of data science and knowledge-based applications. With the advent of the Web and digital technologies, an ever-increasing amount of language data is now available across application areas and industry sectors, including social media, digital archives, company records, etc. The efficient and meaningful exploitation of this data in scientific and commercial innovation is at the core of data science research, employing NLP and machine learning methods as well as semantic technologies based on knowledge graphs. Language data is of increasing importance to machine-learning-based approaches in NLP, linked data and Semantic Web research and applications that depend on linguistic and semantic annotation with lexical, terminological and ontological resources, manual alignment across language or other human-assigned labels. The acquisition, provenance, representation, maintenance, usability, quality as well as legal, organisational and infrastructure aspects of language data are therefore rapidly becoming significant areas of research that are at the focus of the conference.

Knowledge graphs are an active field of research concerned with extracting, integrating, maintaining and using semantic representations of language data in combination with semantically or otherwise structured data, numerical data and multimodal data, among others. Knowledge graph research builds on the exploitation and extension of lexical, terminological and ontological resources, information and knowledge extraction, entity linking, ontology learning, ontology alignment, semantic text similarity, linked data and other Semantic Web technologies. The construction and use of knowledge graphs from language data, possibly and ideally in the context of other types of data, is a further specific focus of the conference.

Furthermore, the conference has also a focus on the emergence of hybrid, neurosymbolic approaches that combine synergistically the great potential of Large Language Models with the explicit semantics contained in knowledge graphs, particularly those containing multilingual data or data from under-resourced languages. A further focus of the conference is the combined use and exploitation of language data and knowledge graphs in data science-based approaches to use cases in industry, including biomedical applications, as well as use cases in humanities and social sciences.

The main conference received 51 submissions, of which 34 were accepted, resulting in an acceptance rate of 66.7%. Accepted works comprised 18 oral presentations (35%) and 16 posters (31%). Each paper was evaluated by three independent reviewers, and the selection process followed a single-blind review format.

This edition of LDK is held in a hybrid format and counts around 70 registered participants, the majority

of them participating onsite in Naples. Jointly with the main conference, we devoted one pre-conference day to host three very interesting workshops. We are publishing the long and short conference papers in a common sub-volume and hosting the proceedings of the workshops in a second one.

Jorge Gracia and Dagmar Gromann
LDK 2025 Conference Chairs

Mehwish Alam and Andon Tchechmedjiev
LDK 2025 Program Committee Chairs

Organizing Committee

Conference Chairs

Jorge Gracia, University of Zaragoza, Spain
Dagmar Gromann, University of Vienna, Austria

Program Chairs

Mehwish Alam, Institut Polytechnique de Paris, France
Andon Tchechmedjiev, Institut Mines-Télécom | EuroMov Digital Health in Motion, France

Workshop Chairs

Katerina Gkirtzou, ILSP "Athena" Research Center, Greece
Slavko Žitnik, University of Ljubljana, Slovenia

Local Organisers

Maria Pia di Buono, University of Naples "L'Orientale", Italy
Johanna Monti, University of Naples "L'Orientale", Italy
Mariapia Battipaglia, University of Naples "L'Orientale", Italy
Argentina Anna Rescigno, University of Naples "L'Orientale", Italy

Publication Chair

Maxim Ionov, University of Zaragoza, Spain

Publicity Chair

Argentina Anna Rescigno, University of Naples "L'Orientale", Italy

Program Committee

Program Committee

Alessandro Adamou, Bibliotheca Hertziana, Max Planck Institute for Art History
Sina Ahmadi, University of Zurich
Mehwish Alam, Telecom Paris, Institut Polytechnique de Paris
Valerio Basile, University of Turin
Carlos Bobed, University of Zaragoza
Francis Bond, Palacký University Olomouc
Federico Boschetti, CNR, Istituto di Linguistica Computazionale "A. Zampolli"
Carmen Brando, EHESS
Eliot Bytyci, University of Prishtina
Sara Carvalho, Universidade de Aveiro
Rute Costa, Universidade Nova de Lisboa
Maria Pia di Buono, University of Naples "L'Orientale"
Milan Dojchinovski, Institute for Applied Informatics and Czech Technical University in Prague
Daniel Fernández-Álvarez, Universidad de Oviedo
Francesca Frontini, CNR, Istituto di Linguistica Computazionale "A. Zampolli"
Katerina Gkirtzou, ILSP "Athena" Research Center
Jorge Gracia, University of Zaragoza
Dagmar Gromann, University of Vienna
Felix Herron, Université Paris Dauphine, PSL
Maxim Ionov, University of Zaragoza
Besim Kabashi, Friedrich-Alexander Universität Erlangen-Nürnberg
Ilan Kernerman, K Dictionaries
Anas Fahad Khan, CNR, Istituto di Linguistica Computazionale "A. Zampolli"
Penny Labropoulou, ILSP "Athena" Research Center
Patricia Martín Chozas, Universidad Politécnica de Madrid
John Philip McCrae, National University of Ireland Galway
Barbara McGillivray, King's College London, University of London
Ana Meštrović, University of Rijeka, Faculty of Informatics and Digital Technologies
Margot Mieskes, University of Applied Sciences, Darmstadt
Elena Montiel Ponsoda, Universidad Politécnica de Madrid
Steven Moran, University of Miami
Hugo Gonçalo Oliveira, Universidade de Coimbra
Ana Ostroški Anić, University of Zagreb
Marco Carlo Passarotti, Università Cattolica del Sacro Cuore
Laurette Pretorius, University of Stellenbosch and University of South Africa
Valeria Quochi, CNR, Istituto di Linguistica Computazionale "A. Zampolli"
Margarida Ramos, Universidade Nova de Lisboa
Paul Rayson, Lancaster University
Georg Rehm, Humboldt-Universität zu Berlin and Deutsches Forschungszentrum für Künstliche Intelligenz
Marko Robnik-Šikonja, University of Ljubljana
Ricardo Rodrigues, Centro de Informática e Sistemas da Universidade de Coimbra and Instituto Politécnico de Coimbra
Anisa Rula, University of Brescia
Harald Sack, Karlsruhe Institute of Technology and FIZ Karlsruhe, Institute for Information Infrastructure

Ana Salgado, CLUNL, Centro de Linguística da Universidade NOVA de Lisboa and Academia das Ciências de Lisboa
Felix Sasaki, SAP SE
Andrea C. Schalley, Karlstad University
Blerina Spahiu, University of Milan, Bicocca
Rachele Sprugnoli, University of Parma
Ranka Stanković, University of Belgrade
Armando Stellato, University of Rome Tor Vergata
Vojtech Svatek, Prague University of Economics and Business
Gilles Sérasset, Université Grenoble Alpes
Andon Tchechmedjiev, IMT Mines Alès
Ciprian-Octavian Truică, University Politehnica of Bucharest
Marieke van Erp, KNAW Humanities Cluster
Vincent Vandeghinste, Instituut voor de Nederlandse Taal, KU Leuven and KU Leuven
Karin Verspoor, Royal Melbourne Institute of Technology
Federica Vezzani, University of Padua
Leon Voukoutis, ILSP "Athena" Research Center
Slavko Žitnik, University of Ljubljana

Keynote Talk

The More You Know: Towards Knowledgeable AI

Gerard de Melo
Hasso Plattner Institute | University of Potsdam

Abstract: The rapid advancement of Generative AI is reshaping the way people search for and acquire knowledge. Yet, despite their impressive capabilities, large language models (LLMs) remain fundamentally unreliable due to their tendency to "hallucinate" — that is, to produce information that is false and not grounded in reality. At the same time, knowledge graphs, while offering structured and reliable facts, also possess important limitations, particularly in terms of their coverage. In light of this, what are viable paths towards more knowledgeable AI systems?

One promising approach lies in extending knowledge graphs by means of machine learning to bridge coverage gaps. This has been the focus of our previous work, including the creation of the Universal WordNet (de Melo and Weikum 2009) and our study on extracting knowledge graphs from language models (Tandon and de Melo 2010).

Another important direction is to better assess and enhance the reliability of LLM outputs. A novel method we explored introduces an explicit I-don't-know marker—the [IDK] token—into the model's vocabulary, paired with a tailored training regimen that encourages the model to select this token when uncertain, rather than generating potentially misleading content (Cohen et al. 2024). We also show how knowledge graphs can contribute to this goal (Cohen et al. 2025).

Finally, a particularly promising avenue is the fusion of LLMs with graph-based knowledge representation. This hybrid approach holds the potential to preserve factual accuracy while improving the transparency and trustworthiness of model outputs (Xian et al. 2019, Bugueño and de Melo 2023, Bugueño et al. 2025).

Together, these directions point toward a future in which AI systems are not only more knowledgeable, but also more reliable and better aligned with human understanding.

Bio: Gerard de Melo is a professor at HPI and the University of Potsdam, where he holds the Chair for AI and Intelligent Systems and leads the corresponding research group. Previously, he was a faculty member at Rutgers University in the US and at Tsinghua University in Beijing, and a post-doc at ICSI/UC Berkeley. Gerard de Melo has published over 200 papers on diverse aspects of AI, receiving a number of Best Paper awards. He served as the General Chair for the AI@HPI Conference and has been featured in the press numerous times.

Keynote Talk

LLMs in Spain: Challenges and Realities

Marta Villergas
Barcelona Supercomputing Center

Abstract: This presentation explores the key challenges and practical realities involved in developing large language models (LLMs) within the Spanish national initiative. It addresses critical topics such as the need for high-performance computing infrastructure, the scarcity and imbalance of data across languages, and issues related to data quality, linguistic and domain coverage, and legal compliance, including data traceability and control.
On the technical side, the talk will cover core components of LLM development—tokenization, pretraining, post-training—as well as evaluation strategies. Particular attention will be paid to the detection and mitigation of bias, ensuring model safety, and integrating ethical principles throughout the development pipeline. The presentation will also highlight derivative models and conclude with reflections on how to build responsible, multilingual AI systems that truly serve diverse linguistic communities.

Bio: Marta Villergas is the Director of the Language Technologies Laboratory at the Barcelona Supercomputing Center (BSC), which is at the forefront of advancing natural language processing (NLP) through pioneering research, development, and the application of high-performance computing (HPC). They specialize in the creation of massive language models and unsupervised learning for less-resourced languages and domains. Endorsed by the Spanish and Catalan governments, the Lab is dedicated to developing vital open-source resources and infrastructure for language technology and artificial intelligence, specifically tailored for the Spanish and Catalan languages. Marta Villergas has been engaged in various EU-funded international projects and am committed to promoting the transfer of our technological breakthroughs to industry and society at large.

Keynote Talk

Do Large Language Models Understand Word Meanings?

Roberto Navigli
Sapienza University of Rome

Abstract: The ability to interpret word meanings in context is a core yet underexplored challenge for Large Language Models (LLMs). While these models demonstrate remarkable linguistic fluency, the extent to which they genuinely grasp word semantics remains an open question. In this talk, we investigate the disambiguation capabilities of state-of-the-art instruction-tuned LLMs, benchmarking their performance against specialized systems designed for Word Sense Disambiguation (WSD). We also examine lexical ambiguity as a persistent challenge in Machine Translation (MT), particularly when dealing with rare or context-dependent word senses. Through an in-depth error analysis of both disambiguation and translation tasks, we reveal systematic weaknesses in LLMs, shedding light on the fundamental challenges they face in semantic interpretation. Furthermore, we show the limitations of standard evaluation metrics in capturing disambiguation performance, reinforcing the need for more targeted evaluation frameworks. By presenting dedicated testbeds, we introduce more effective ways to assess lexical understanding both within and across languages, and highlight the gap between the impressive fluency of LLMs and their actual semantic comprehension of language.

Bio: Roberto Navigli is a professor of Natural Language Processing at the Sapienza University of Rome, where he leads the Sapienza NLP Group. He has received two ERC grants on multilingual semantics, highlighted among the 15 projects through which the ERC has transformed science. He has received several prizes, including two Artificial Intelligence Journal prominent paper awards and several outstanding/best paper awards from ACL. He leads the Italian Minerva LLM Project — the first LLM pre-trained in Italian — and is the Scientific Director and co-founder of Babelscape, a successful deep-tech company developing next-generation multilingual NLU and NLG. He is a Fellow of ACL, AAAI, EurAI and ELLIS, and serves as General Chair of ACL 2025.

Table of Contents