3rd Conference on Language, Data and Knowledge

LDK 2021, September 1–3, 2021, Zaragoza, Spain

Edited by

Dagmar Gromann
Gilles Sérasset
Thierry Declerck
John P. McCrae
Jorge Gracia
Julia Bosque-Gil
Fernando Bobillo
Barbara Heinisch

Editors

Dagmar Gromann
University of Vienna, Austria
dagmar.gromann@gmail.com

Thierry Declerck
DFKI GmbH, Germany
declerck@dfki.de

Jorge Gracia
University of Zaragoza, Spain
jogracia@unizar.es

Fernando Bobillo
University of Zaragoza, Spain

Gilles Sérasset
Université Grenoble Alpes, France
gilles.serasset@imag.fr

John P. McCrae
National University of Ireland Galway, Ireland
john.mccrae@insight-centre.org

Julia Bosque-Gil
University of Zaragoza, Spain

Barbara Heinisch
University of Vienna, Austria
barbara.heinisch@univie.ac.at

ACM Classification 2012
Computing methodologies → Natural language processing; Computing methodologies → Knowledge representation and reasoning

ISBN 978-3-95977-199-3

Published online and open access by
Schloss Dagstuhl – Leibniz-Zentrum für Informatik GmbH, Dagstuhl Publishing, Saarbrücken/Wadern, Germany. Online available at https://www.dagstuhl.de/dagpub/978-3-95977-199-3.

Publication date
August, 2021

Bibliographic information published by the Deutsche Nationalbibliothek
The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at https://portal.dnb.de.

License
This work is licensed under a Creative Commons Attribution 4.0 International license (CC-BY 4.0): https://creativecommons.org/licenses/by/4.0/legalcode.
In brief, this license authorizes each and everybody to share (to copy, distribute and transmit) the work under the following conditions, without impairing or restricting the authors' moral rights:
- Attribution: The work must be attributed to its authors.

The copyright is retained by the corresponding authors.

Digital Object Identifier: 10.4230/OASIcs.LDK.2021.0

ISBN 978-3-95977-199-3 **ISSN 1868-8969** **https://www.dagstuhl.de/oasics**

OASIcs – OpenAccess Series in Informatics

OASIcs is a series of high-quality conference proceedings across all fields in informatics. OASIcs volumes are published according to the principle of Open Access, i.e., they are available online and free of charge.

Editorial Board

- Daniel Cremers (TU München, Germany)
- Barbara Hammer (Universität Bielefeld, Germany)
- Marc Langheinrich (Università della Svizzera Italiana – Lugano, Switzerland)
- Dorothea Wagner (*Editor-in-Chief*, Karlsruher Institut für Technologie, Germany)

ISSN 1868-8969

https://www.dagstuhl.de/oasics

Contents

Preface
Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch 0:ix

Organizing Committee
..... 0:xi

Scientific Advisory Committee
..... 0:xiii

Program Committee
..... 0:xv

Invited Talks

The JeuxDeMots Project
Mathieu Lafourcade 1:1–1:1

A Smell is Worth a Thousand Words: Olfactory Information Extraction and Semantic Processing in a Multilingual Perspective
Sara Tonelli 2:1–2:1

Free/Open-Source Machine Translation for the Low-Resource Languages of Spain
Mikel L. Forcada 3:1–3:1

Crazy New Ideas

A Computational Simulation of Children's Language Acquisition
Ben Ambridge 4:1–4:3

Get! Mimetypes! Right!
Christian Chiarcos 5:1–5:4

Mind the Gap: Language Data, Their Producers, and the Scientific Process
Tobias Weber 6:1–6:9

Language Data

Representing the Under-Represented: a Dataset of Post-Colonial, and Migrant Writers
Marco Antonio Stranisci, Viviana Patti, and Rossana Damiano 7:1–7:14

Plenary Debates of the Parliament of Finland as Linked Open Data and in Parla-CLARIN Markup
Laura Sinikallio, Senka Drobac, Minna Tamper, Rafael Leal, Mikko Koho, Jouni Tuominen, Matti La Mela, and Eero Hyvönen 8:1–8:17

Towards a Corpus of Historical German Plays with Emotion Annotations
Thomas Schmidt, Katrin Dennerlein, and Christian Wolff 9:1–9:11

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Enriching a Lexical Resource for French Verbs with Aspectual Information
Anna Kupść, Pauline Haas, Rafael Marín, and Antonio Balvet 10:1–10:12

Annotation of Fine-Grained Geographical Entities in German Texts
Julián Moreno-Schneider, Melina Plakidis, and Georg Rehm 11:1–11:8

Supporting the Annotation Experience Through CorEx and Word Mover's Distance
Stefania Pecòre 12:1–12:15

A Twitter Corpus and Lexicon for Abusive Speech Detection in Serbian
Danka Jokić, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih 13:1–13:17

Knowledge Graphs

Bias in Knowledge Graphs – An Empirical Study with Movie Recommendation and Different Language Editions of DBpedia
Michael Matthias Voit and Heiko Paulheim 14:1–14:13

Enriching Word Embeddings with Food Knowledge for Ingredient Retrieval
Álvaro Mendes Samagaio, Henrique Lopes Cardoso, and David Ribeiro 15:1–15:15

TatWordNet: A Linguistic Linked Open Data-Integrated WordNet Resource for Tatar
Alexander Kirillovich, Marat Shaekhov, Alfiya Galieva, Olga Nevzorova, Dmitry Ilvovsky, and Natalia Loukachevitch 16:1–16:12

Explainable Zero-Shot Topic Extraction Using a Common-Sense Knowledge Graph
Ismail Harrando and Raphaël Troncy 17:1–17:15

Relevance Feedback Search Based on Automatic Annotation and Classification of Texts
Rafael Leal, Joonas Kesäniemi, Mikko Koho, and Eero Hyvönen 18:1–18:15

Automatic Construction of Knowledge Graphs from Text and Structured Data: A Preliminary Literature Review
Maraim Masoud, Bianca Pereira, John McCrae, and Paul Buitelaar 19:1–19:9

An Ontology for CoNLL-RDF: Formal Data Structures for TSV Formats in Language Technology
Christian Chiarcos, Maxim Ionov, Luis Glaser, and Christian Fäth 20:1–20:14

On the Utility of Word Embeddings for Enriching OpenWordNet-PT
Hugo Gonçalo Oliveira, Fredson Silva de Souza Aguiar, and Alexandre Rademaker 21:1–21:13

Applications for Language, Data and Knowledge

Towards Learning Terminological Concept Systems from Multilingual Natural Language Text
Lennart Wachowiak, Christian Lang, Barbara Heinisch, and Dagmar Gromann .. 22:1–22:18

Encoder-Attention-Based Automatic Term Recognition (EA-ATR)
Sampritha H. Manjunath and John P. McCrae 23:1–23:13

Universal Dependencies for Multilingual Open Information Extraction
Massinissa Atmani and Mathieu Lafourcade 24:1–24:15

Inconsistency Detection in Job Postings
Joana Urbano, Miguel Couto, Gil Rocha, and Henrique Lopes Cardoso 25:1–25:16

A Workbench for Corpus Linguistic Discourse Analysis
Julia Krasselt, Matthias Fluor, Klaus Rothenhäusler, and Philipp Dreesen 26:1–26:9

APiCS-Ligt: Towards Semantic Enrichment of Interlinear Glossed Text
Maxim Ionov 27:1–27:8

Introducing the NLU Showroom: A NLU Demonstrator for the German Language
Dennis Wegener, Sven Giesselbach, Niclas Doll, and Heike Horstmann 28:1–28:9

AAA4LLL – Acquisition, Annotation, Augmentation for Lively Language Learning
Bartholomäus Wloka and Werner Winiwarter 29:1–29:15

Improving Intent Detection Accuracy Through Token Level Labeling
Michał Lew, Aleksander Obuchowski, and Monika Kutyła 30:1–30:11

Towards Scope Detection in Textual Requirements
Ole Magnus Holter and Basil Ell 31:1–31:15

Discrepancies Between Database- and Pragmatically Driven NLG: Insights from QUD-Based Annotations
Christoph Hesse, Maurice Langner, Anton Benz, and Ralf Klabunde 32:1–32:9

Bridging the Gap Between Ontology and Lexicon via Class-Specific Association Rules Mined from a Loosely-Parallel Text-Data Corpus
Basil Ell, Mohammad Fazleh Elahi, and Philipp Cimiano 33:1–33:21

Use Cases in Language, Data and Knowledge

HISTORIAE, History of Socio-Cultural Transformation as Linguistic Data Science. A Humanities Use Case
Florentina Armaselu, Elena-Simona Apostol, Anas Fahad Khan, Chaya Liebeskind, Barbara McGillivray, Ciprian-Octavian Truică, and Giedrė Valūnaitė Oleškevičienė 34:1–34:13

An Automatic Partitioning of Gutenberg.org Texts
Davide Picca and Cyrille Gay-Crosier 35:1–35:9

A Data Augmentation Approach for Sign-Language-To-Text Translation In-The-Wild
Fabrizio Nunnari, Cristina España-Bonet, and Eleftherios Avramidis 36:1–36:8

A Review and Cluster Analysis of German Polarity Resources for Sentiment Analysis
Bettina M. J. Kern, Andreas Baumann, Thomas E. Kolb, Katharina Sekanina, Klaus Hofmann, Tanja Wissik, and Julia Neidhardt 37:1–37:17

Exploring Causal Relationships Among Emotional and Topical Trajectories in Political Text Data
Andreas Baumann, Klaus Hofmann, Bettina Kern, Anna Marakasova, Julia Neidhardt, and Tanja Wissik 38:1–38:8

Calculating Argument Diversity in Online Threads
Cedric Waterschoot, Antal van den Bosch, and Ernst van den Hemel 39:1–39:9

Linking Discourse Marker Inventories
Christian Chiarcos and Maxim Ionov 40:1–40:15

Tackling Domain-Specific Winograd Schemas with Knowledge-Based Reasoning and Machine Learning
Suk Joon Hong and Brandon Bennett 41:1–41:13

Preface

This volume presents the proceedings of the 3rd Conference on Language, Data and Knowledge (LDK 2021) held in Zaragoza, Spain from September 1–3, 2021. Language, Data and Knowledge 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), and it has been supported by an international Scientific Committee of leading researchers in Natural Language Processing, Linked Data and Semantic Web, Language Resources and Digital Humanities. This initial conference was successfully continued in the second edition of LDK in Leipzig, Germany in 2019, organized by the Institut für Angewandte Informatik (InfAI) and co-organized by the Insight Centre for Data Analytics and the Applied Computational Linguistics Lab (ACoLi).

This third edition of the LDK conference is hosted by the University of Zaragoza in Zaragoza, Spain. Major support was provided by the NexusLinguarum COST Action CA18209 "European network for Web-centred linguistic data science", the Prêt-à-LLOD project funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No 825182, and the University of Zaragoza.

As a biennial event, LDK aims at bringing together researchers from across disciplines concerned with the acquisition, curation and use of 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, organizational and infrastructure aspects of language data are therefore rapidly becoming major areas of research that are at the focus of the conference.

Knowledge graphs is an active field of research concerned with the extraction, integration, maintenance and use of 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.

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.

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

In total, 71 papers were submitted and reviewed by 69 reviewers. Typically, at least 3 reviews per paper resulted in 38 accepted papers for oral and poster presentations. As a novel feature, LDK 2021 had a special track for "Crazy New Ideas", that is, short abstracts that provide the occasion to present challenging research ideas that have not yet been fully explored or that the researchers would like to see in ten years from now. This category was decidedly aimed at research creativity and sparking interesting novel discussions within the LDK community. Three such crazy new ideas could be accepted for LDK and for publication.

Organizing Committee

Conference Chairs

John P. McCrae (National University of Ireland Galway, Ireland)
Thierry Declerck (DFKI GmbH, Germany)

Local Organizers

Julia Bosque Gil (University of Zaragoza, Spain)
Fernando Bobillo (University of Zaragoza, Spain)
Jorge Gracia (University of Zaragoza, Spain)

Program Chairs

Dagmar Gromann (University of Vienna, Austria)
Gilles Sérasset (Université Grenoble Alpes, France)

Workshop Chairs

Sara Carvalho (Universidade de Aveiro, Portugal)
Renato Rocha Souza (Austrian Academy of Sciences, Austria)

Proceedings Chair

Barbara Heinisch (University of Vienna, Austria)

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Scientific Advisory Committee

John P. McCrae (National University of Ireland Galway, Ireland)
Paul Buitelaar (National University of Ireland Galway, Ireland)
Christian Chiarcos (Goethe-University Frankfurt, Germany)
Tatjana Gornostaja (Tilde, Latvia)
Philipp Cimiano (Bielefeld University, Germany)
Gerard de Melo (Rutgers University, USA)
Francis Bond (Nanyang Technological University, Singapore)
Thierry Declerck (DFKI GmbH, Germany)
Franciska de Jong (CLARIN ERIC, the Netherlands)
Karin Verspoor (University of Melbourne, Australia)
Edward Curry (National University of Ireland Galway, Ireland)
Jorge Gracia (University of Zaragoza, Spain)
Nancy Ide (Vassar College, USA)
Milan Dojchinovski (InfAI @ Leipzig University, Germany / CTU in Prague, Czech Republic)

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch

OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Program Committee

Alessandro Adamou (The Open University)

Nathalie Aussenac-Gilles (IRIT CNRS)

Denilson Barbosa (University of Alberta)

Valerio Basile (University of Turin)

Pierpaolo Basile (University of Bari)

Martin Benjamin (Kamusi Project International)

Michael Bloodgood (The College of New Jersey)

Julia Bosque-Gil (Universidad de Zaragoza)

Paul Buitelaar (NUI Galway)

Harry Bunt (Tilburg University)

Aljoscha Burchardt (DFKI GmbH)

Eliot Bytyçi (University of Prishtina)

Nicoletta Calzolari (Istituto di Linguistica Computazionale – CNR)

Philipp Cimiano (Bielefeld University)

Gerard de Melo (HPI, University of Potsdam)

Thierry Declerck (DFKI GmbH)

Milan Dojchinovski (Czech Technical University in Prague)

Patrick Ernst (Amazon)

Maria Eskevich (CLARIN ERIC)

Luis Espinosa-Anke (Cardiff University)

Thierry Fontenelle (European Investment Bank)

Francesca Frontini (Istituto di Linguistica Computazionale A.Zampolli – CNR – Pisa)

Debanjan Ghosh (Educational Testing Service)

Hugo Gonçalo Oliveira (University of Coimbra)

Jeff Good (University at Buffalo)

Eero Hyvönen (Aalto University and University of Helsinki (HELDIG))

Nancy Ide (Vassar College)

Sepehr Janghorbani (Rutgers University)

Besim Kabashi (Friedrich-Alexander-Universität Erlangen-Nürnberg)

Te Taka Keegan (Waikato University)

Ilan Kernerman (K Dictionaries)

Dimitris Kontokostas (University of Leipzig)

Maria Koutraki (L3S Research Center, Leibniz University of Hannover)

Udo Kruschwitz (University of Regensburg)

Chaya Liebeskind (Jerusalem College of Technology, Lev Academic Center)

John P. McCrae (National University of Ireland Galway)

Margot Mieskes (University of Applied Sciences, Darmstadt)

Steven Moran (University of Neuchâtel)

Diego Moussallem (Paderborn University)

Alessandro Oltramari (Bosch Research and Technology Center)

Petya Osenova (Sofia University and IICT-BAS)

Bolette Pedersen (University of Copenhagen)

Laurette Pretorius (School of Interdisciplinary Research and Graduate Studies, University of South Africa)

Gábor Prószéky (MorphoLogic & PPKE)

Francesca Quattri (The Hong Kong Polytechnic University)

Alexandre Rademaker (IBM Research Brazil and EMAp/FGV)

3rd Conference on Language, Data and Knowledge (LDK 2021).
Editors: Dagmar Gromann, Gilles Sérasset, Thierry Declerck, John P. McCrae, Jorge Gracia, Julia Bosque-Gil, Fernando Bobillo, and Barbara Heinisch
OpenAccess Series in Informatics
OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

Simon Razniewski (Max Planck Institute for Informatics)

Georg Rehm (DFKI GmbH)

Nils Reiter (Institute of Natural Language Processing, Stuttgart University)

Steffen Remus (University of Hamburg)

Laurent Romary (INRIA & HUB-ISDL)

Mike Rosner (University of Malta)

Marco Rospocher (University of Verona)

Felix Sasaki (Cornelsen Verlag GmbH & TH Brandenburg)

Andrea Schalley (Karlstad University)

Max Silberztein (Université de Franche-Comté)

Steffen Staab (IPVS, Universität Stuttgart, DE and WAIS, University of Southampton, UK)

Armando Stellato (University of Rome, Tor Vergata)

Stan Szpakowicz (University of Ottawa)

Liling Tan (Nanyang Technological University)

Ciprian-Octavian Truica (Aarhus University)

Andrius Utka (Vytautas Magnus University)

Giedre Valunaite Oleskeviciene (Mykolas Romeris University)

Marieke van Erp (KNAW Humanities Cluster)

Marc Verhagen (Brandeis University)

Karin Verspoor (RMIT University)

Piek Vossen (Vrije Universiteit Amsterdam)

Qian Yang (Duke University)

Ziqi Zhang (Sheffield University)