LREC 2020 Workshop
Language Resources and Evaluation Conference
11–16 May 2020

7th Workshop on Linked Data in Linguistics
(LDL-2020)

PROCEEDINGS

Editors:
Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, Jorge Gracia

Proceedings of the LREC 2020
7th Workshop on Linked Data in Linguistics
(LDL-2020)

Edited by:
Maxim Ionov, John P. McCrae, Christian Chiarcos, Thierry Declerck, Julia Bosque-Gil, and Jorge Gracia

The 7th Workshop on Linked Data in Linguistics was supported by the European Union's Horizon 2020 research and innovation program under grant agreement no. 825182 through the Prêt-à-LLOD project, the German Ministry for Education and Research (BMBF) through the project Linked Open Dictionaries (LiODi, 2015-2020) and the COST Action CA18209, NexusLinguarum: European network for Web-centred linguistic data science.

ISBN: 979-10-95546-36-8
EAN: 9791095546368

For more information:
European Language Resources Association (ELRA)
9 rue des Cordelières
75013, Paris
France
http://www.elra.info
Email: lrec@elda.org

7th Workshop on Linked Data in Linguistics (LDL-2020). Building tools and infrastructures

Past years have seen a growing interest in the application of knowledge graphs and Semantic Web technologies to language resources, and their publication as linked data on the Web. As of today, a large amount of language resources were either converted or created natively as linked data on the basis of data models specifically designed for the representation of linguistic content. Examples are wordnets, dictionaries, corpora, culminating in the emergence of a Linguistic Linked Open Data (LLOD) cloud (`http://linguistic-lod.org/`).

Since its establishment in 2012, the Linked Data in Linguistics (LDL) workshop series has become the major forum for presenting, discussing and disseminating technologies, vocabularies, resources and experiences regarding the application of semantic technologies and the Linked Open Data (LOD) paradigm to language resources in order to facilitate their visibility, accessibility, interoperability, reusability, enrichment, combined evaluation and integration. The LDL workshops contribute to the discussion, dissemination and establishment of community standards that drive this development, most notably the OntoLex-lemon model for lexical resources, as well as standards for other types of language resources still under development.

The workshop series is organized by Open Linguistics, founded 2010 as a Working Group of the Open Knowledge Foundation[1] with close involvement of related communities, such as W3C Community Groups, and international research projects. It takes a general focus on LOD-based resources, vocabularies, infrastructures and technologies as means for managing, improving and using language resources on the Web. As technology and resources increasingly converge towards a LOD-based ecosystem, this year we particularly encouraged submissions on Linked-Data Aware Tools and Services and Linked Language Resources Infrastructure, i.e. managing, curating and applying LLOD technologies and resources in a reliable and reproducible way for the needs of linguistics, NLP and digital humanities.

After ten years of community work, a critical mass of LLOD resources is already in place, yet, there is still a need to develop a robust ecosystem of tools that consume linguistic linked data. Recently started research networks and European projects are working in the direction of building sustainable infrastructures around LRs, with linked data as one of the core technologies. LDL-2020 is thus supported by the COST Action "European network for Web-centred linguistic data science" (NexusLinguarum) and two Horizon 2020 projects, the European Lexicographic Infrastructure (ELEXIS), and Prêt-à-LLOD, which focuses on providing an infrastructure for linguistic data to be ready to use by state-of-the-art technologies.

With a focus on building tools and applications, the *7th Workshop on Linked Data in Linguistics (LDL-2020)* was organized in conjunction with the 12th Language Resource and Evaluation Conference (LREC-2020). We received a total of 23 submissions out of which 12 were accepted (acceptance rate 52%). Due to Covid-19, LDL-2020 was not taking place as a physical meeting, but as a virtual event[2]. Presentations of the accepted papers were organized in three groups with four presentations each, on modelling, applications and lexicography, respectively.

[1] `https://groups.google.com/forum/#!forum/open-linguistics`

[2] Details and the program are available at `http://ldl2020.linguistic-lod.org/program.html`

Modelling

In *Towards an ontology based on Hallig-Wartburg's Begriffssystem for Historical Linguistic Linked Data* Tittel et al. compare two strategies for the LOD modelling of a conceptual system that is used in historical lexicography and lexicology, based on SKOS and OWL, respectively, and with examples from medieval Gascon and Italian.

In *Transforming the Cologne Digital Sanskrit Dictionaries into Ontolex-Lemon*, Mondaca and Rau evaluate two strategies for transforming TEI/XML data into OntoLex-Lemon, the enrichment of TEI XML with RDFa data, and a native RDF modelling. This evaluation tackles an important issue for applications in Digital Humanities as the TEI does not provide commonly accepted specifications for interfacing traditional XML-based workflows and Linked Open Data technologies.

In *Representing Temporal Information in Lexical Linked Data Resources*, Khan describes recent developments on his extension of the OntoLex-Lemon vocabulary with diachronic lexical information with examples from the Oxford English Dictionary and an etymological dictionary.

In *From Linguistic Descriptions to Language Profiles*, Shafqat Mumtaz Virk et al. introduce the concept of language profiles as structured representations of various types of knowledge about a natural language, they describe how to semi-automatically construct such data from descriptive documents and they develop a language profile of an example language.

Applications and Infrastructures

While overarching linked data-based infrastructures are only emerging, numerous applications of this technology are being reported.

With *Terme-à-LLOD: Simplifying the Conversion and Hosting of Terminological Resources as Linked Data*, Maria Pia di Buono et al. simplify the transformation and publication of terminology data by virtualization: A preconfigured virtual image of a server can thus be used to simplify installation of transformation and hosting services for terminological resources as linked data.

Frank Abromeit et al. introduce *Annohub – Annotation Metadata for Linked Data Applications*, a dataset and a portal that provides metadata about annotation and language identification for annotated language resources available on the web. Annohub builds on metadata repositories to identify language resources, on automated routines for classifying languages and annotation schemes, a broad range of transformers for various corpus formalisms and human curation for quality assurance.

Salgado et al. address *Challenges of Word Sense Alignment* for Portuguese Language Resources and report on a comparative study between the Portuguese Academy of Sciences Dictionary and the Dicionário Aberto. Word sense alignment involves searching for matching senses within dictionary entries of different lexical resources and linking them, implemented here by means of Semantic Web technologies.

In *A Lime-Flavored REST API for Alignment Services*, Fiorelli and Stellato describe a REST API to enable the participation of downstream alignment services an orchestration framework for ontology alignment. Using explicit metadata about the input ontologies, other resources and the task itself, a report is produced that summarizes characteristics and alignment strategies. For the lexical content of the input ontologies and external language resources, the report uses the Lime module of the OntoLex-Lemon model.

Lexicography

Abgaz describes on-going work on *Using OntoLex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects*, comprising two main components, a questionnaire with details about questions, collectors, paper slips etc., and a lexical dataset which contains lexical entries (answers) collected in response to the questions. The paper describes how the original TEI/XML format is transformed into Linguistic Linked Open Data to produce a lexicon for Bavarian Dialects.

With Linguistic Linked (Open) Data and, especially, the OntoLex vocabulary now being widely adapted throughout lexicography, there is a demand for tools, both for exploiting linked lexical data and for creating a user-friendly access to it. In *Involving Lexicographers in the LLOD Cloud with LexO, an Easy-to-use Editor of Lemon Lexical Resources*, Bellandi and Giovannetti describe LexO, a collaborative web editor of OntoLex-Lemon resources.

As for tools for lexicography, Gun Woo Lee et al. describe *Supervised Hypernymy Detection in Spanish through Order Embeddings*, based on a hypernymy dataset for Spanish built from WordNet and the use of pretrained word vectors as input.

Finally, Nielsen reports on *Lexemes in Wikidata*, i.e., the way that Wikidata records data about lexemes, senses and lexical forms and exposes them as Linguistic Linked Open Data and the growth and development of this data set since its first establishment in 2018.

Organizers:

Maxim Ionov, Goethe University Frankfurt (Germany)
John P. McCrae, National University of Ireland, Galway (Ireland)
Christian Chiarcos, Goethe University Frankfurt (Germany)
Thierry Declerck, DFKI GmbH (Germany) and ACDH-ÖAW (Austria)
Julia Bosque-Gil, University of Zaragoza (Spain)
Jorge Gracia, University of Zaragoza (Spain)

Program Committee:

Paul Buitelaar, Insight (Ireland)
Steve Cassidy, Macquarie University (Australia)
Philipp Cimiano, University of Bielefeld (Germany)
Gerard de Melo, Rutgers University (USA)
Francesca Frontini, Université Paul-Valéry (France)
Jeff Good, University at Buffalo (USA)
Dagmar Gromann, Vienna University (Austria)
Yoshihiko Hayashi, Osaka University, Waseda University (Japan)
Fahad Khan, ILC-CNR (Italy)
Bettina Klimek, University of Leipzig (Germany)
Elena Montiel-Ponsoda, Universidad Politécnica de Madrid (Spain)
Steve Moran, Universität Zürich (Switzerland)
Roberto Navigli, "La Sapienza" Università di Roma (Italy)
Sebastian Nordhoff, Language Science Press Berlin (Germany)
Petya Osenova, IICT-BAS (Bulgaria)
Antonio Pareja-Lora, Universidad Complutense Madrid (Spain)
Laurent Romary, INRIA (France)
Felix Sasaki, Cornelsen Verlag GmbH Berlin (Germany)
Andrea Schalley, Karlstad University (Sweden)
Gilles Sérasset, University Grenoble Alpes (France)
Armando Stellato, University of Rome, Tor Vergata (Italy)
Marieke van Erp, KNAW Humanities Cluster (The Netherlands)
Piek Vossen, Vrije Universiteit Amsterdam (The Netherlands)

Table of Contents