EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School

Milan Dojchinovski, Julia Bosque Gil, Jorge Gracia, Ranka Stanković

EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School | Milan Dojchinovski, Julia Bosque Gil, Jorge Gracia, Ranka Stanković | Infotheca | 2021 | |

10.18485/infotheca.2021.21.1.7

http://dr.rgf.bg.ac.rs/s/repo/item/0005143

EUROLAN 2021: Introduction to Linked Data for Linguistics Online Training School

UDC 81: 37.018.53
DOI 10.18485/infotheca.2021.21.1.7

ABSTRACT: The first training school organized by the *NexusLinguarum* COST Action was held on February 8-12, 2021 and was aimed at students, academics, and practitioners wishing to learn the basics of Linguistic Data Science. During the training school, the participants were introduced to a wide range of topics: from Semantic Web, RDF and ontologies, to modeling and querying linguistic data with state-of-the-art ontology models and tools. The training school was organized under the umbrella of the EUROLAN series of summer schools and was hosted virtually (online) by several institutions: the Romanian Academy, the Research Institute for Artificial Intelligence in Bucharest and the Institute of Computer Science in Iași, as well as the "Alexandru Ioan Cuza" University of Iași, Romania. The training school was attended by 82 participants.

KEYWORDS: linguistic data science, linked data for linguistics, language data, NexusLinguarum, COST action, EUROLAN, training school.

PAPER SUBMITTED: 30 June 2021
PAPER ACCEPTED: 13 July 2021

Milan Dojchinovski
milan.dojchinovski@fit.cvut.cz
CTU in Prague
Prag, Czech Republic
InfAI at Leipzig University
Leipzig, Germany

Julia Bosque Gil
jbosque@unizar.es
Jorge Gracia
jogracia@unizar.es
Aragon Institute of Engineering Research (I3A)
University of Zaragoza
Zaragoza, Spain

Ranka Stanković
ranka.stankovic@rgf.bg.ac.rs
University of Belgrade
Faculty of Mining and Geology
Serbia

1 Introduction

NexusLinguarum - European network for Web-centered linguistic data science, COST action CA18029[1] - was launched at the end of October 2019. The goal of the *NexusLinguarum* action is to promote the study of linguistic data science, for which the construction of an ecosystem of multilingual

1. NexusLinguarum

and semantically interoperable linguistic data is required. Training schools are one of the means for reaching this goal, and therefore the *NexusLinguarum* core team organized the Introduction to Linked Data for Linguistics online training school[2] that took place from February 8 to 12, 2021. The training school was aimed at promoting and teaching the basics of linguistic data science and the related technologies to people from the academia and the industry. It was organized under the umbrella of the EUROLAN series of summer schools, which was established in 1993 and covers topics that are particularly relevant to the fields of computational linguistics and natural language processing (NLP). The goal of this 15th EUROLAN School was to bring together scholars, teachers and students of linguistics, NLP and information technology to discuss the principles and best practices for representing, publishing and linking linguistic data and the issues that constitute the building blocks in the envisioned multilingual and interoperable web-oriented ecosystem. The present contribution summarises the organisation, content and results of this training school and is based on Deliverable D1.1[3] of the Action.

2 The training school program

The training school has been developed for newcomers as well as for those already having basic knowledge in the fields covered. The school provided a comprehensive introduction to the methodologies for representing linguistic resources using semantic web technologies, together with the means to extract knowledge from language resources and exploit it using semantic web query languages and reasoning capabilities. The topics addressed in the school were the following:

- Semantic Web and Linked Data[4] (Berners-Lee et al. 2006);
- Ontologies: RDF (Resource description framework), RDF Schema (Resource Description Framework Schema, variously abbreviated as RDFS, RDF(S), RDF-S, or RDF/S), Web Ontology Language (OWL),[5] etc.);
- SPARQL query language- a semantic query language for databases able to retrieve and manipulate data stored in the RDF format;

2. EUROLAN
3. Deliverable D1.1
4. Introducing Linked Data and the Semantic Web
5. OWL

- Metadata: DCAT (Data Catalog Vocabulary),[6] VOID (RDF Schema vocabulary for expressing metadata about RDF datasets, etc.);
- RDF transformation and validation; (Cimiano et al. 2020)
- Linguistic linked data; (Chiarcos et al. 2013)
- Lemon-OntoLex[7] (McCrae et al. 2017; Declerck, Tiberius, and Wandl-Vogt 2017; Stanković et al. 2018)
- Linguistic linked data generation; (Cimiano et al. 2020)
- Corpora and linked data; (Chiarcos 2012)
- Linguistic annotations; (Fäth et al. 2020)
- NLP Interchange Format; (Hellmann et al. 2013)
- Tools and applications of linguistic linked data. (Declerck et al. 2020)

The first day started with an opening session and a brief introduction to Linguistic Linked Data (LLD), followed by an introduction to Linked Data and RDF dedicated sessions. The second day covered topics related to ontologies, including modelling knowledge with ontologies, OWL and SKOS[8] knowledge representation languages, reasoning of knowledge, and a hands-on session using the Protégé[9] ontology editor. The third day was dedicated to the topics related to representing and querying lexical data with dedicated sessions on the OntoLex-Lemon model and the SPARQL querying language. The fourth day included sessions which gave an overview of other linguistic and metadata vocabularies and the VocBench platform[10] (Stellato et al. 2020) modelling linguistic datasets. In the afternoon, an online social event was organized where the participants could remotely see the beauty of the Romanian culture, traditions and nature. The fifth day comprised three parallel sessions on different topics:

(i.) LLD Generation/Transformation and Linking,
(ii.) Annotations (NIF, Web Annotation) (Hellmann et al. 2013), and
(iii.) OntoLex extensions: *vartrans* for representing translations and term variants (based on the *lemon* translation module, (Gracia et al. 2014)), *lexicog*[11] – lexicography module (Bosque-Gil, Gracia, and Montiel-

6. Data Catalog Vocabulary (DCAT) - Version 2
7. Lemon - Lexicon Model for Ontologies; Lexicon Model for Ontologies: Community Report, 10 May 2016
8. SKOS Simple Knowledge Organization System - home page
9. Protégé
10. VocBench: A Collaborative Management System for OWL ontologies, SKOS(/XL) thesauri, OntoLex-lemon lexicons and generic RDF datasets
11. The OntoLex Lemon Lexicography Module

Ponsoda 2017), *FrAC*[12] – frequency, attestation and corpus Information (Chiarcos et al. 2020).

Finally, the training school ended with a closing session where an ontology of participants, lecturers and organizers was presented, illustrating many of the representation mechanisms explained throughout the week.

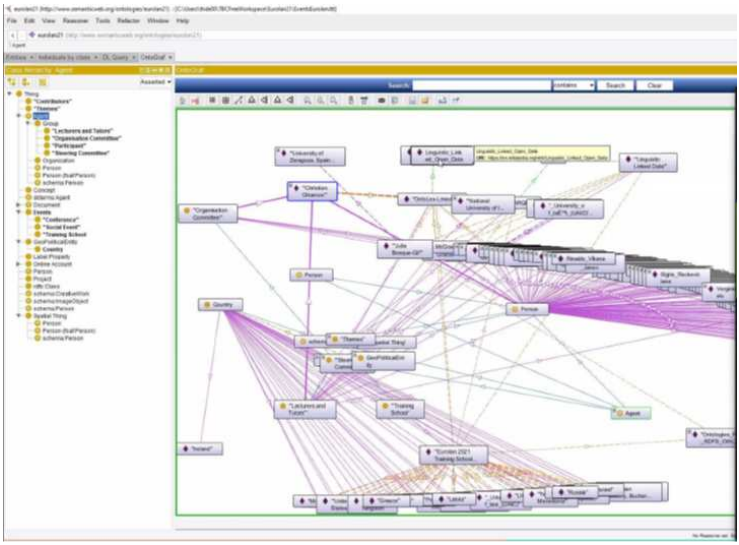

Figure 1. Ontology of the training school.

.

Each of the organized sessions was accompanied by a hands-on session and an exercise session. During the hands-on session, the lecturers proposed an exercise and offered a step-by-step walk-through for the participants to understand the methodology leading towards the solution. They also introduced the basic technology needed. Then, during the exercise session, the participants were asked to work on a particular task like the cases presented during the hands-on session, thereby becoming familiar with the technology introduced in a practical setting. As these sessions were graded in terms of complexity, starting with the basic notions, and building on to present more specific topics in a detailed fashion on the last day, the participants had

12. FrAC – Frequency, Attestation and Corpus Information - Ontology-Lexica Community Group

a chance to acquire a solid foundation before moving onto more complex sessions. The official program of the school is available online.[13]

As a follow up, the JeRTeh[14] Language Resources and Technologies Society set up a local installation of VocBench[15] and, apart from JeRTeh members, it was used by students and teachers of the Intelligent Systems PhD program[16] at the University of Belgradefor the subjects Knowledge representation and Semantic web. The Lemon-OntoLex Frac module was used for representation of the entries from the lexicon used for abusive speech detection with attestations from the Twitter corpus with annotation of abusive spans (Jokić et al. 2021).

3 Organization

Due to the COVID-19 pandemic and current travel restrictions in Europe and beyond, the training school was held online. Following on the almost three decades long tradition of EUROLAN, which is known for academic program excellence and camaraderie among professors and students, a range of virtual activities were carried out in addition to holding online classes, with the aim of providing cultural experiences and discoveries, in addition to closer interaction. Attendance was online and free of charge, requiring pre-registration.

All the sessions were hosted using a videoconferencing platform. For the hands-on sessions, several breakout (virtual) rooms were made available where the participants could work on the assignment in smaller groups. To encourage participants to ask questions and get in touch with each other, the organizers set up a Slack[17] channel, as a collaboration hub where lecturers and participants could clarify any doubts. The total number of participants was 82, 52 female and 30 male, including 4 participants from Serbia.

Various types of materials were generated for the training school, including presentations (slides)[18] and exercises[19] accompanied by code and data

13. School Program
14. Jerteh
15. VocBench installation
16. Intelligent Systems PhD Program
17. Slack, The School channel
18. Presentations
19. Exercises

examples.[20] All the materials were published online and made available for free.

4 Summary

The training school provided valuable knowledge and trained many computer scientists and linguists on how to work with and benefit from linguistic linked data. This was the first training school organized by the *NexusLinguarum* COST Action as one of a series of training events that are planned to take place. It aimed to serve as an introduction to the topic of linguistic data science and build the basis for the audience necessary for attending future training schools on more advanced topics for the duration of the COST Action. All the materials created during the training school are publicly available and can be further used by the community. During the closing session, the organizers provided participants with a survey form to gather feedback on both organizational and academic aspects of the school. The results have shown that the disciplines of the humanities/linguistics/lexicography had a higher representation among participants than computer science, and that the school was well-focused, well-balanced topic-wise and well organized. Theory sessions, tutoring, and the opportunities to learn were very highly evaluated. On the other hand, due to the virtual mode, there is still room for improvement in practical sessions, social event organization and opportunities to network. The knowledge and skills acquired there will improve the development of Serbian linguistic resources and help to publish more resources as linguistic linked data.

Acknowledgment

This paper is supported by the COST Action CA18209 - *NexusLinguarum* "European Network for Web-centred Linguistic Data Science".

References

Berners-Lee, Tim, Yuhsin Chen, Lydia Chilton, Dan Connolly, Ruth Dhanaraj, James Hollenbach, Adam Lerer, and David Sheets. 2006. "Tabulator: Exploring and analyzing linked data on the semantic web." In *Proceedings of the 3rd international semantic web user interaction workshop,* 2006:159. Athens, Georgia.

20. Supporting material

Bosque-Gil, Julia, Jorge Gracia, and Elena Montiel-Ponsoda. 2017. "Towards a Module for Lexicography in OntoLex." In *LDK Workshops,* 74–84.

Chiarcos, Christian. 2012. "Interoperability of corpora and annotations." In *Linked Data in Linguistics,* 161–179. Springer.

Chiarcos, Christian, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. 2020. "Modelling frequency and attestations for ontolex-lemon." In *Proceedings of the 2020 Globalex Workshop on Linked Lexicography,* 1–9.

Chiarcos, Christian, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. "Towards open data for linguistics: Linguistic linked data." In *New Trends of Research in Ontologies and Lexical Resources,* 7–25. Springer.

Cimiano, Philipp, Christian Chiarcos, John P McCrae, and Jorge Gracia. 2020. "Converting language resources into linked data." In *Linguistic Linked Data,* 163–180. Springer.

Declerck, Thierry, John Philip McCrae, Matthias Hartung, Jorge Gracia, Christian Chiarcos, Elena Montiel-Ponsoda, Philipp Cimiano, Artem Revenko, Roser Sauri, Deirdre Lee, et al. 2020. "Recent developments for the linguistic linked open data infrastructure." In *Proceedings of the 12th LREC,* 5660–5667.

Declerck, Thierry, Carole Tiberius, and Eveline Wandl-Vogt. 2017. "Encoding lexicographic data in lemon: Lessons learned." In *Proceedings of the LDK workshops: OntoLex, TIAD and Challenges for Wordnets. CEURS,* vol. 8.

Fäth, Christian, Christian Chiarcos, Björn Ebbrecht, and Maxim Ionov. 2020. "Fintan-flexible, integrated transformation and annotation engineering." In *Proceedings of the 12th LREC,* 7212–7221.

Gracia, Jorge, Elena Montiel-Ponsoda, Daniel Vila-Suero, and Guadalupe Aguado-De-Cea. 2014. "Enabling Language Resources to Expose Translations as Linked Data on the Web." In *Proceedings of the 9th LREC,* edited by Nicoletta Calzolari (Conference Chair) et al. Reykjavik, Iceland: European Language Resources Association (ELRA), May. ISBN: 978-2-9517408-8-4.

Hellmann, Sebastian, Jens Lehmann, Sören Auer, and Martin Brümmer. 2013. "Integrating NLP using linked data." In *International Semantic Web Conference,* 98–113. Springer.

Jokić, Danka, Ranka Stanković, Cvetana Krstev, and Branislava Šandrih. 2021. "A Twitter Corpus and lexicon for abusive speech detection in Serbian." In *Proceedings of the 2021 Language, Data and Knowledge (LDK), 1-3 September in Zaragoza, Spain.*

McCrae, John P, Julia Bosque-Gil, Jorge Gracia, Paul Buitelaar, and Philipp Cimiano. 2017. "The Ontolex-Lemon model: development and applications." In *Proceedings of eLex 2017 conference,* 19–21.

Stanković, Ranka, Cvetana Krstev, Biljana Lazić, and Mihailo Škorić. 2018. "Electronic dictionaries-from file system to lemon based lexical database." In *Proceedings of the 11th LREC - W23 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science (LDL-2018), LREC 2018, Miyazaki, Japan, May 7-12, 2018,* 48–56.

Stellato, Armando, Manuel Fiorelli, Andrea Turbati, Tiziano Lorenzetti, Willem Van Gemert, Denis Dechandon, Christine Laaboudi-Spoiden, Anikó Gerencsér, Anne Waniart, Eugeniu Costetchi, et al. 2020. "VocBench 3: A collaborative Semantic Web editor for ontologies, thesauri and lexicons." *Semantic Web* 11 (5): 855–881.